# *Estimación de las relaciones de parentesco realizadas entre medio-hermanos con distintos métodos*

*Tesis para optar al título de Magister de la Universidad de Buenos Aires*
*Área Biometría y Mejoramiento*

**Carolina Andrea García Baccino**
Ing. Agr. – FAUBA - 2012

Lugar de trabajo: Cátedra de Mejoramiento Genético Animal. Facultad de Agronomía.
Universidad de Buenos Aires


Escuela para Graduados Ing. Agr. Alberto Soriano
Facultad de Agronomía – Universidad de Buenos Aires

## COMITÉ CONSEJERO

**Director de tesis**
**Rodolfo Juan Carlos Cantet**
Ingeniero Agrónomo (Universidad de Buenos Aires)
MSc (Montana State University)
MSc (University of Illinois)
PhD (University of Illinois)

**Consejero de Estudios**
**Sebastián Munilla Leguizamón**
Ingeniero Agrónomo (Universidad de Buenos Aires)
Doctor en Ciencias Agropecuarias (Universidad de Buenos Aires)

## JURADO DE TESIS

**Director de tesis**
**Rodolfo Juan Carlos Cantet**
Ingeniero Agrónomo (Universidad de Buenos Aires)
MSc (Montana State University)
MSc (University of Illinois)
PhD (University of Illinois)

**Jurado**
**Laura Puhl**
Ingeniera Agrónoma (Universidad de Buenos Aires)
MSc en Biometría (Universidad de Buenos Aires)
Doctora en Ciencias Agropecuarias (Universidad de Buenos Aires)

**Jurado**
**Cecilia Bruno**
Ingeniera Agrónoma (Universidad de Buenos Aires)
MSc en Ciencias Agropecuarias (Universidad Nacional de Córdoba)
MSc en Estadística Aplicada (Universidad Nacional de Córdoba)
Doctora Ciencias Agropecuarias (Universidad Nacional de Córdoba)

**Jurado**
**Guillermo Giovambattista**
Licenciado en Biología (Universidad Nacional de La Plata, Argentina)
Doctor en Ciencias Naturales (Universidad Nacional de La Plata, Argentina)

Fecha de defensa de la tesis: 22 de mayo de 2017

*Puedes llegar a cualquier parte, siempre que andes lo suficiente.*
(Lewis Carroll**,** **"**Alicia en el País de las Maravillas"**)**

A mis padres por darme las herramientas y la libertad
para construir y recorrer mi camino.

A mis abuelos y Matías por mostrarme que no importan
los obstáculos, siempre se puede llegar a cualquier parte.

# AGRADECIMIENTOS

*Me gustaría agradecer a un grupo de personas e instituciones que contribuyeron enormemente a que esta tesis llegara a concretarse.*

*Como primera medida me gustaría agradecer a los miembros de mi comité: Dr. Rodolfo Juan Carlos Cantet (Fito) y Dr. Sebastián Munilla Leguizamón. Fito, gracias por todos y cada uno de sus sabios consejos, sin ellos nada de esto hubiese sido posible. Muchas gracias por enseñarme desde el ejemplo y sobre todo por darme la libertad y confianza para trabajar en lo que más me gusta y el espacio para enfrentar nuevos desafíos. Al Dr. Sebastián Munilla Leguizamón. Gracias Sebas por guiarme e impulsarme incansablemente para lograr este objetivo, por todos los consejos y apoyo que me permitieron seguir adelante siempre, sobre todo en momentos decisivos.*

*A Zulma y Andrés quienes me recibieron gentilmente, aconsejaron, impulsaron y apoyaron para lograr este objetivo.*

*Al Dr. Elsen por leer mi trabajo y hacerme sugerencias muy valiosas.*

*A los miembros del jurado por tomarse el trabajo de leer esta tesis y hacer valiosos comentarios para mejorarla.*

*A la Facultad de Agronomía de la Universidad de Buenos Aires y a la cátedra de Mejoramiento Genético Animal por el apoyo institucional.*

*A CONICET por darme la posibilidad, a través de una beca de iniciar y culminar esta tesis.*

*A INRA Toulouse e INP Toulouse por financiar mi estancia en Francia para terminar este trabajo.*

*Al Dr. Juan Pedro Steibel por cederme el uso de los datos de la población experimental de cerdos y por sus valiosos comentarios.*

*A Marito por esos ratos de mates, charlas y ánimo para seguir hacia adelante.*

*A Susana y Amelia, por todo su cariño y alegría.*

*A todo el personal de la Escuela para graduados entre los que me gustaría destacar a María Del Carmen Fabrizio por su dedicación, trabajo incansable y pasión por enseñar.*

*A mis compañeros y amigos de Biometría: María Isabel, Camilo y Pablo por acompañarme en todo el proceso de cursada y aprendizaje.*

*A mis compañeros y amigos de la Cátedra de Mejoramiento Genético Animal: Sebas, Nati, Andresito, Juan, Joselito, Yeni, Mati, Dani, Belcy, Esteban y Martín. Gracias por todos sus consejos, apoyo y alegría cotidiana.*

*A Valeria Schindler, Laura Pruzzo, Mónica Santos Cristal por todas las enseñanzas al inicio de mi camino en la cátedra. Una mención especial para Anita Birchmeier por transmitirme su alegría y pasión al trabajar y por todos sus sabios consejos.*

*A mis amigos por acompañarme, aconsejarme y entenderme todos estos años.*

*A mis padres, Claudia y Daniel, por su apoyo incondicional y sabios consejos.*

*A mis hermanos por acompañarme siempre.*

*A Matías, por estar siempre, entenderme y acompañarme. Por ser mi apoyo en momentos decisivos.*

*A Cocó quien me enseñó a cuestionarme todo y a trabajar incansablemente hasta alcanzar mis objetivos.*

*A todos muchas gracias.*

*Declaro que el material incluido en esta tesis es, a mi mejor saber y entender, original producto de mi propio trabajo (salvo en la medida en que se identifique explícitamente las contribuciones de otros), y que este material no lo he presentado, en forma parcial o total, como una tesis en ésta u otra institución.*

Carolina Andrea García Baccino

# PUBLICACIONES DERIVADAS DE LA TESIS

- García Baccino, C. A., Munilla, S., Legarra, A., Vitezica, Z.G., Forneris, N.S., Ernst, C.W., Bates, R.O., Raney, N., Steibel, J.P. y Cantet, R. J. C. 2016. Estimates of the actual relationship between half-sibs in a pig population. *Journal of Animal Breeding and Genetics*. 1 – 10

# ÍNDICE GENERAL

# ÍNDICE DE CUADROS

# ÍNDICE DE FIGURAS

**FIGURA** **página**

# ABREVIATURAS

| | |
|---|---|
| **CV** | Coeficiente de Variación |
| **DS** | Desvío Estándar |
| **EHW** | Equilibrio Hardy-Weinberg |
| **GWR** | Relación de parentesco realizada |
| **HMM** | Modelos Ocultos de Markov (del inglés *Hidden Markov Models*) |
| **IBD** | Identidad por descendencia (del inglés *Identity by Descent*) |
| **IBS** | Identidad por estado (del inglés *Identity by State*) |
| **LD** | Desequilibrio de ligamiento (del inglés *Linkage Disequilibrium*) |
| **LE** | Equilibrio de ligamiento (del inglés *Linkage Equilibrium*) |
| **MAF** | Alelo en menor proporción (del inglés *Minor allele frequency*) |
| **MM** | Modelo Markoviano |
| **MME** | Ecuaciones de Modelo Mixto (del inglés *Mixed Model Equations*) |
| **PIBD** | Proporción de genoma idéntico por descendencia |
| **QTL** | Locus de un carácter cuantitativo |
| **SNP** | Polimorfismos genéticos de un solo nucleótido (del inglés *Single Nucleotide Polymorphysm*) |

# RESUMEN

## *Estimación de las relaciones de parentesco realizadas entre medio-hermanos con distintos métodos*

Las relaciones genómicas calculadas empleando la información de marcadores moleculares miden la proporción real del genoma compartida de manera "idéntica por descendencia" (PIBD), entre dos individuos. Existen diversos métodos para estimar las relaciones de parentesco realizadas (GWR). En este trabajo se compararon los desempeños de cuatro métodos sobre la base de las distribuciones empíricas de las GWR estimadas, con 6704 parejas de medio-hermanos de una población de cerdos cruza. Tres de esas metodologías emplean información genómica y genealógica (siguiendo un enfoque IBD) para estimar las GWR y el restante utiliza solo datos genómicos (IBS). Dentro del primer grupo, uno de los métodos que estima las probabilidades de transmisión de segmentos cromosómicos de padres a hijos y toma en cuenta los bloques de ligamiento, mostró la distribución empírica de las GWR más cercana a la teórica esperada, seguido de aquél que emplea un modelo oculto de Markov y tiene en cuenta el desequilibrio de ligamiento (LD). El tercer método dentro de este grupo, no toma en cuenta el LD y generó distribuciones empíricas alejadas de la esperada, especialmente en términos del desvío standard (SD). Los cambios en la cantidad de información genómica afectaron sólo a los dos últimos métodos. Por otra parte, el procedimiento que sigue un enfoque IBS generó distribuciones empíricas de GWR cuya media y SD fueron superiores a los valores teóricos esperados, y el valor de la media se vio afectado al emplear diferentes cantidades de datos genómicos. Los resultados obtenidos reflejan un desempeño cercano al teórico esperado por parte de los métodos IBD que combinan genealogía e información molecular, destacándose aquellos que consideran el LD. Concretamente, el método que estima las probabilidades de transmisión de padres a hijos fue el que produjo las distribuciones empíricas de las GWR entre medio-hermanos más cercanas a la teórica esperada, incluso ante variaciones en la cantidad de información genómica disponible.

**Palabras clave:** Identidad por descendencia, relaciones de parentesco realizadas, métodos de estimación, selección genómica.

# ABSTRACT

## *Estimates of the actual relationship between half-sibs using different methods*

Genomic relationships based on markers capture the actual instead of the expected (based on pedigree) proportion of genome shared identical by descent (IBD). Several methods exist to estimate genomic relationships. In this research we compare four such methods that were tested looking at the empirical distribution of the estimated relationships across 6,704 pairs of half sibs from a crossbred pig population. Three of those methods use both genomic and pedigree information (IBD approach) and the remaining one only considers genomic information (IBS). Within the first group, one method that estimates the transmission probabilities of chromosomal segments from parents to offspring taking into account the linkage groups displayed a mean and standard deviation (SD) in close agreement with the expected ones. A method based on a hidden Markov model that takes into account linkage disequilibrium (LD) came second when comparing its estimates with the expected ones. The third method from this group, does not consider LD and showed the smallest empirical SD. Changes in the amount of the genomic information available affected the last two methods. On the other hand, the method following an IBS approach displayed a mean and a SD higher than the expected ones and the mean was sensible to changes in the amount of marker information. The results show that the methods that follow an IBD approach performed closer to theoretical values, especially those that consider LD. Within this group, the method that estimates the transmission probabilities from parents to offspring was the one that performed closer to theoretical values, even when changing the amount of genomic information.



**Key words:** Identity by descent; actual relationships; estimation methods; genomic selection.

# Capítulo 1. *Introducción*

# Capítulo 1

# Introducción

Cuantificar el grado de parentesco entre individuos es esencial en muchas áreas de la genética. En el ámbito del mejoramiento genético animal es de vital importancia para la predicción del mérito genético y la estimación de las heredabilidades de los caracteres. Con tal fin, se utilizó tradicionalmente un enfoque basado en la identidad por descendencia (IBD) que permite calcular las relaciones de parentesco esperadas entre dos individuos. Dos genes en un locus determinado son idénticos por descendencia cuando ambos son copias idénticas del mismo gen ancestral (Malecot, 1969). Más específicamente, estas medidas de parentesco toman un rol central en la predicción de los valores de cría, a través de la matriz ***A*** de relaciones aditivas (Henderson, 1984). Sus elementos son iguales a dos veces la probabilidad de que dos genes homólogos en dos individuos distintos sean IBD, y se calculan empleando información genealógica. La información de marcadores moleculares permite evaluar además la real proporción del genoma IBD compartida entre dos individuos (Guo, 1995) y calcular las relación de parentesco observada o "realizada" entre ellos que, de ahora en más, denotaremos como GWR. No obstante, existe cierta variabilidad de GWR entre pares de individuos que poseen la misma relación esperada (Hill, 1993 a; Hill y Weir, 2011). Esto se debe a la estocasticidad de procesos como la segregación mendeliana y la recombinación durante la gametogénesis, procesos que determinan cuáles segmentos cromosómicos recibirá cada individuo de sus ancestros. Este proceso estocástico genera la variabilidad responsable de que, por ejemplo, dos parejas de medio-hermanos no compartan necesariamente la misma proporción de genoma IBD, como se asume al construir la matriz ***A***. Esta es la causa por la cual las relaciones de parentesco esperadas reflejan de modo incompleto a las GWR (Speed y Balding, 2015).

En los últimos años surgió un modo alternativo para predecir el mérito genético empleando la información de paneles densos de marcadores compuestos por polimorfismos genéticos de un solo nucleótido (SNP). Una estrategia comúnmente empleada para predecir el valor de cría consiste en estimar el efecto de cada marcador SNP de una población animal, para un carácter de interés con herencia poligénica (Meuwissen *et al.*, 2001). Una vez obtenidas, las estimaciones para cada marcador son sumadas con el objeto de calcular un valor de cría genómico para cada uno de los animales evaluados. Otra alternativa consiste en emplear un modelo lineal equivalente (bajo el supuesto de normalidad) en el que se modifican las ecuaciones de modelo mixto (MME) reemplazando la matriz ***A*** de relaciones aditivas esperadas por un matriz ***G*** de relaciones de parentesco genómicas observadas. Esta matriz genómica puede ser computada siguiendo un enfoque de identidad por estado (IBS), empleando sólo información molecular (VanRaden, 2008). En este caso, la matriz ***G*** es el resultado de considerar el contenido génico en cada loci, valor que permite obtener una estimación de la proporción de genoma compartido IBD (VanRaden 2007; VanRaden 2008; Toro *et al.*, 2011). Alternativamente, ***G*** puede calcularse siguiendo un enfoque IBD condicional a la información de los marcadores y de la genealogía. A tal

efecto se propusieron una gran cantidad de métodos contrastantes en términos de los supuestos sobre los cuales trabajan y de la demanda computacional que conllevan. Entre ellos se encuentran los métodos propuestos por Guo (1994), Elsen *et al.* (2009) y Han y Abney (2011). Por lo tanto, los objetivos generales de la tesis son: 1. Evaluar y comparar el desempeño de cuatro métodos contrastantes para estimar GWR entre pares de medio-hermanos de una base de datos real. 2. Evaluar el impacto de variar la cantidad y calidad de la información genómica sobre las estimaciones obtenidas para cada uno de los cuatro métodos. Esto último pone de manifiesto la capacidad de cada uno de ellos para capturar las GWR frente a cambios en las bases de datos genómicos.

El documento está organizado en seis capítulos siendo el primero esta introducción. El siguiente contiene una revisión bibliográfica que detalla los antecedentes de los métodos de estimación de GWR (Capítulo 2). Dicha sección constituye el marco teórico de referencia donde se describe la gran variedad de métodos para estimar GWR disponibles a la fecha, cada uno de ellos con una base teórica y/o algorítmica distinta. Además, se explican detalladamente las cuatro metodologías evaluadas, destacando los supuestos sobre los cuales trabajan y la demanda computacional que conllevan. En el Capítulo 3 se describe la base de datos analizada y los procedimientos utilizados para su edición con el objetivo de generar dos conjuntos de datos genómicas diferentes, en términos de cantidad y calidad de información contenida. También se describe el modo en que fueron implementados cada uno de los métodos y los programas empleados para realizar los cálculos. Los resultados obtenidos con cada uno de los métodos se presentan en el Capítulo 4. El capítulo 5 contiene una discusión general de los resultados y, finalmente, el capítulo 6 enumera las conclusiones generales de la tesis.

# Capítulo 2. *Revisión Bibliográfica*

# Capítulo 2

# Revisión Bibliográfica

## 2.1. Estimación de las relaciones de parentesco realizadas.

Dos genes en un locus dado son idénticos por descendencia (IBD) si ambos son copias idénticas del mismo gen ancestral (Malecot, 1969). El enfoque tradicional basado en la identidad por descendencia permite calcular las relaciones de parentesco esperadas entre dos individuos. En dicho caso el cálculo de la probabilidad de observar IBD entre los individuos *X* e *Y* $\left(P\left(X \equiv Y\right)\right)$ es condicional a la información genealógica y se calcula bajo el supuesto de apareamiento aleatorio y equilibrio Hardy-Weinberg (HWE). Por lo tanto, se establece implícitamente que dicha probabilidad es igual en cualquier punto del genoma. En la actualidad y gracias al advenimiento de las técnicas moleculares que identifican SNPs (del inglés "*Single Nucleotide Polymorphysm*", polimorfismos genéticos de un sólo nucleótido) es posible estimar las relaciones de parentesco realizadas u observadas (GWR). Contar con la información adicional de los genotipos permite calcular qué proporción del genoma comparten dos individuos emparentados. Esta se puede calcular de dos maneras distintas: i) siguiendo un enfoque IBS (identidad en estado) en el que sólo se toma en cuenta la información de los marcadores moleculares, sin considerar la genealogía, tal como lo hace el método propuesto por VanRaden (2008), o ii) calculando la proporción de genoma idéntico por descendencia (PIBD) compartido entre dos individuos tomando en cuenta la información genealógica (*P*) y la de los marcadores moleculares (*M*), del siguiente modo $PIBD = P\left(X \equiv Y \mid P, M\right)$. Se ha desarrollado un gran número de metodologías para estimar las relaciones de parentesco realizadas u observadas. Los mayores avances en este campo se dieron inicialmente en el ámbito de la genética humana, con metodologías sencillas para obtener aproximaciones de la proporción de genoma compartido entre dos individuos en regiones puntuales del genoma sin considerar aspectos tales como la consanguinidad o la falta de independencia entre marcadores moleculares. Estos métodos fueron evolucionando, al involucrar modelos de complejidad creciente y algoritmos más eficientes para lograr estimaciones cada vez más precisas de PIBD sobre el genoma completo.

Gagnon *et al.* (2005) destacaron la relevancia de trabajar con el enfoque IBD como modo natural y lógico de evaluar el parecido entre individuos. Dentro de la gran diversidad de métodos disponibles para estimar PIBD, se pueden hallar los algoritmos de Lander y Green (1987) y Fishelson y Geiger (2002) que permiten el cálculo exacto de genoma IBD compartido, pero son inviables para el cómputo cuando los pedigríes son extensos y complejos. Asimismo, Goldgar (1990) fue el primero en presentar un método "multipunto" para estimar PIBD entre pares de hermanos en una región cromosómica específica comprendida entre dos marcadores (microsatélites en éste caso). La noción de "multipunto" refiere a que la metodología emplea simultáneamente varios marcadores para obtener

información más precisa respecto del número de eventos de recombinación que se producen en segmentos cromosómicos específicos. Los modelos "multipunto" permiten modelar fidedignamente el proceso de herencia de material genético puesto que, como indican Guo (1995) y Thompson (2013), el genoma se transmite en segmentos y no por puntos. Esta metodología fue extendida por el mismo Guo (1994, 1995, 1996) quien realizó un análisis teórico más profundo sobre PIBD. Sus estudios se fundamentaron en el trabajo de Donnelly (1983), quien propuso utilizar una cadena de Markov en dos estados para modelar el proceso genético que se da en cada cromosoma. Guo (1995) extendió el concepto a un grupo de individuos y presentó algoritmos para calcular las probabilidades de interés. Posteriormente, Gagnon *et al.* (2005) fueron los primeros en llevar a cabo un estudio empírico del tema, aplicando las metodologías descriptas a bases de datos reales con información de microsatélites para estimar PIBD entre hermanos, tomando en cuenta toda la extensión del genoma.

Existen algoritmos alternativos como los de Cadenas de Markov vía muestreo de Monte Carlo (*Markov Chain Monte Carlo*, Thompson, 2000). Estos métodos permiten aproximar PIBD, pero presentan dificultades al trabajar con genealogías complejas que cuentan con gran número de generaciones de individuos sin datos y pueden acarrear problemas de convergencia (Fernandez *et al.*, 2001). Posteriormente, Leutenegger *et al.* (2003) propusieron un modelo de Markov de tiempo continuo para modelar el proceso IBD entre dos cromosomas dentro de un mismo individuo, siendo el tiempo la distancia genética a lo largo del cromosoma. Dado que dicho enfoque no permite el análisis de PIBD, Thompson (2007) lo extendió para considerar cuatro cromosomas en dos individuos. Abney (2008) propuso un método computacionalmente eficiente que permite trabajar con un gran número de individuos y genealogías extensas, sobre la base de un modelo oculto de Markov (HMM). Si bien este algoritmo presenta varias ventajas a nivel computacional, las estimaciones de PIBD son de un sólo "punto" (del inglés "*single point*"), es decir, son condicionales a la información genotípica de *sólo un locus*, siendo esta la principal falencia del método propuesto.

Tradicionalmente las estimaciones de PIBD han sido calculadas bajo el supuesto que las frecuencias conjuntas entre marcadores son iguales al producto de las frecuencias marginales de los marcadores, es decir independencia estadística. Este supuesto se empleó durante el período que se extiende entre el trabajo de Thompson (1975) hasta Anderson y Weir (2007). Dicho supuesto es insostenible en la actualidad, dado que los paneles de marcadores de alta densidad han mostrado una importante asociación entre SNPs (Daly *et al.*, 2001). El *desequilibrio de ligamiento* (LD, del inglés de "*Linkage disequilibrium*") o *desequilibrio gamético*, es la asociación no aleatoria de alelos ubicados en dos o más loci. El LD se manifiesta cuando el producto de las frecuencias alélicas marginales es distinto de la frecuencia genotípica conjunta, aun cuando las primeras estén en equilibrio Hardy-Weinberg en cada locus en particular. Como consecuencia del LD, los genes o haplotipos pueden segregar sin cambios en su conformación, de una generación a la siguiente. Este es el motivo por el cual es necesario tener en cuenta el efecto del LD al estimar las probabilidades de IBD; al modelarlo adecuadamente es posible reducir la cantidad de falsos positivos en la detección de IBD, y aumentar la potencia del método empleado al poder detectar segmentos IBD de mucho menor tamaño (Browning y Browning, 2010). Existen métodos más recientes como los propuestos por Elsen *et al.* (2009), Li *et al.* (2010) y Han y

Abney (2011) que toman en cuenta el LD a la hora de estimar las PIBD. El primero se caracteriza por ser altamente eficiente dado que permite calcular las probabilidades de transmisión de segmentos cromosómicos de padres a hijos, utilizando la mínima información necesaria dentro de cada bloque de ligamiento. Estos tres métodos producen estimaciones de PIBD "multipunto". Tanto el método de Han y Abney (2011) como el de Li *et al.* (2010) emplean un HMM con modificaciones que permiten considerar la *dependencia* existente entre genotipos producto del LD. El primero posibilita además, considerar la consanguinidad, mientras que el segundo toma en cuenta el IBD poblacional, también conocido como "Background IBD", para captar el LD entre SNPs. El IBD poblacional describe el parentesco histórico entre individuos que va más allá de la información que se puede obtener a partir de la genealogía conocida.

En las siguientes secciones se describirán cuatro métodos diferentes que permiten calcular las GWR. Este subconjunto de cuatro métodos busca representar la vasta variedad de metodologías disponibles en la actualidad para tal fin. Cada una de ellas se diferencia en aspectos tales como la utilización de información de genealogía y/o marcadores moleculares, la consideración o no del LD y la consanguinidad, los supuestos específicos sobre los cuales trabajan y el modelo estadístico que emplean.

## 2.2. El método de Guo (1994)

El modelo propuesto por Guo (1994) permite calcular la PIBD compartida entre dos individuos medio-hermanos. El hecho de limitar los cálculos a un sólo tipo de relación de parentesco en la que ambos individuos considerados comparten un progenitor en común, permite modelar el proceso de transmisión de material genético mediante una cadena de Markov (MM) con dos estados posibles, donde el parámetro de tiempo refiere a la longitud de mapa de cada segmento. Esta metodología está desarrollada sobre la base de la teoría presentada por Donnelly (1983), quien mostró que el proceso de "crossing-over" en un pedigrí es asimilable a una cadena de Markov de tiempo continuo cuyos estados son los vértices de un hiper-cubo (Stefanov, 2002). Donnelly (1983) introdujo el concepto de pedigrí cromosómico con el objetivo de modelar el proceso de transmisión de ADN por segmentos utilizando un proceso estocástico continuo. En dicho pedigrí, cada cromosoma del individuo de interés resulta del proceso de recombinación de los cromosomas parentales de uno de los padres, es decir de aquellos segmentos provenientes de los abuelos (maternos o paternos) del individuo bajo análisis (Bickeböller y Thompson, 1996). Empleando la notación 0 para indicar proveniencia del cromosoma materno y 1 para la proveniencia paterna, el proceso de "crossing-over" se comporta de modo "markoviano" entre ambos estados. En el caso de medio-hermanos y empleando SNPs, se asigna entonces 1 al alelos del SNP en una posición particular del genoma cuando proviene del cromosoma del abuelo paterno del individuo, o es igual a 0 cuando corresponde al cromosoma originado en la abuela paterna del individuo. En la Figura 2.1 se presenta un breve ejemplo para un par de medio-hermanos paternos.

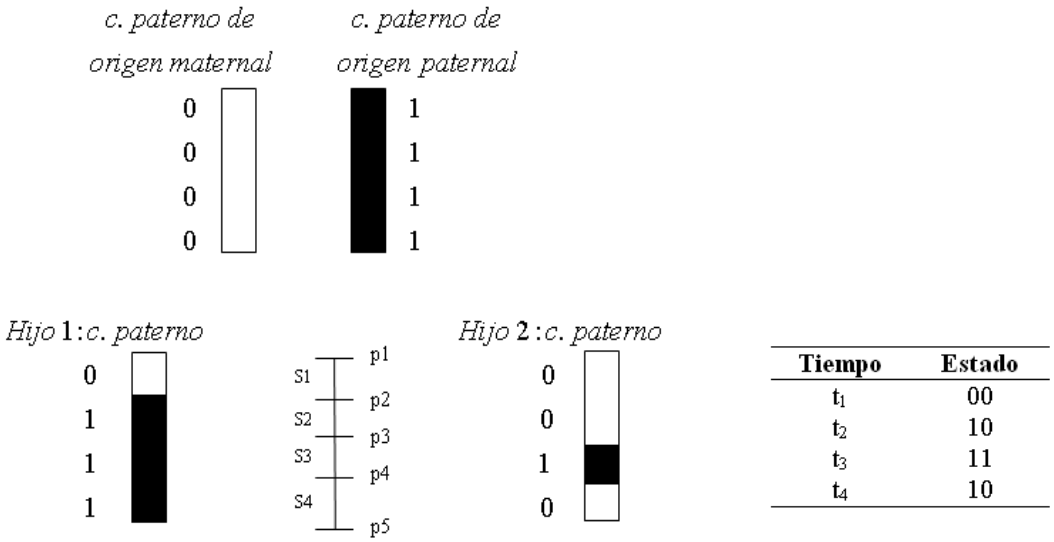



**Figura 2.1.** Ejemplo de un pedigrí cromosómico para medio-hermanos paternos.

La Figura 2.1 muestra que ambos individuos poseen el cromosoma materno, aquél proveniente de la abuela paterna, en el segmento S1. El valor de estado en ese segmento es 00. En la posición p2 se detecta un "crossing-over" en el individuo 1, por lo que el estado en el segmento S2 pasa a ser 10. Al presentarse otro "crossing-over" en la posición p3 del indiviuo 2, el valor de estado en el segmento S3 se transforma en 11. Finalmente, la "caminata aleatoria" (random walk) alcanza el estado 10 nuevamente luego de un nuevo "crossing-over" en posición p4. Cabe destacar que la Figura 2.1 tiene como objetivo ilustrar cómo funciona el método y la representación de los estados, motivo por el cual se presentó un caso en el que el número de "crossing-over" es superior al esperado en casos reales sólo una meiosis.

El método presentado por Guo (1994) emplea los siguientes supuestos: 1) el modelo de recombinación es el de Haldane (1919), modelando los "crossing-overs" a lo largo del cromosoma mediante un proceso estocástico de Poisson; 2) no existe diferencia alguna en la longitud de mapa entre ambos sexos; 3) ausencia de alteraciones del genoma, como ser mutaciones, translocaciones, conversiones, deleciones o inserciones; 4) la población base está constituida por individuos no emparentados (ausencia de consanguinidad); 5) las fases parentales son conocidas. Se entiende por *fase* parental a aquella combinación de alelos que un individuo recibió de sus padres. Esta metodología considera todas las configuraciones posibles que pueden darse con los alelos de los marcadores al tomar un segmento genómico definido entre dos SNPs. Para el caso de parejas de medio-hermanos los casos posibles que pueden darse en cada segmento a lo largo de todo el genoma son 16 y son detallados en el Cuadro 2.1. Cabe destacar que los primeros siete casos fueron derivados por Goldgar (1990) y posteriormente Guo (1994) los extendió para incluir las situaciones de mayor grado de incertidumbre; es decir, aquellas donde los marcadores parentales no son completamente informativos, o cuando existen genotipos faltantes. Cada uno de los casos presentados en el Cuadro 2.1 se asocia a una expresión que permite el cálculo tanto la esperanza como de la varianza de PIBD para el segmento considerado. Dichas expresiones

dependen de dos parámetros: λ y θ. El primero corresponde a la distancia de mapa entre los dos marcadores (*i* y *j*) que determinan al segmento considerado, distancia que es referida en Morgans. El segundo parámetro se calcula del siguiente modo: $\theta = 0.5\left(1 - e^{-2\lambda}\right)$. Además, en la segunda a la quinta columna del Cuadro 2.1, *i* toma valor 0 si el alelo del marcador proviene por origen materno (abuela), mientras que es igual a 1 cuando proviene del lado paterno (abuelo). A modo de ejemplo, en el caso 5 se consideran las cuatro situaciones posibles en las que los alelos son IBD para sólo uno de ambos loci considerados. En esta situación, ambos individuos comparten el alelo paterno o materno en uno de los marcadores y, en el otro locus los alelos de ambos individuos no son IBD (uno de ellos es de origen materno mientras que el restante es paterno). De este modo, al avanzar segmento a segmento por el genoma de la pareja de medio-hermanos, es necesario primero evaluar qué caso se observa en cada segmento, para posteriormente calcular PIBD para cada uno de ellos, según la expresión correspondiente a la situación observada. Una vez computadas las probabilidades asociadas con cada segmento cromosómico se procede a calcular un valor global de PIBD sobre todo el genoma para los medio-hermanos *x* e *y*, ponderando las probabilidades de cada segmento (*i* hasta *N*) según la longitud del mismo ($l_i$), tal como lo indica la siguiente expresión

$$\text{PIBD}_{x,y\,\text{global}} = \frac{\sum_{i=1}^{N} l_i \ \text{PIBD}_i}{L} \qquad [1]$$

siendo *L* la longitud de todo el genoma autosómico, más específicamente $L = \sum_{i=1}^{N} l_i$.

**Cuadro 2.1**. Esperanza y varianza de la PIBD para cada uno de los casos propuestos por Guo (1994). Los índices $i_1$ y $j_1$ corresponden a dos marcadores consecutivos evaluados en el individuo 1 y, análogamente, $i_2$ y $j_2$ son los índices de los mismos marcadores en el animal 2. Los individuos considerados son medio-hermanos. El valor λ es la distancia de mapa entre los marcadores *i* y *j* medida en Morgans y $\theta = 0,5\left(1 - e^{-2\lambda}\right)$. Desde la segunda a la quinta columna, $i = 0$ cuando el alelo proviene del lado materno e $i = 1$ cuando proviene por herencia paterna.

| Caso | $i_1$ | $i_2$ | $j_1$ | $j_2$ | E (PIBD) | Var (PIBD) |
|---|---|---|---|---|---|---|
| **1** | $i$ | $i$ | $i$ | $I$ | $\frac{2(1-\theta)^2\lambda + \theta(1-\theta) + \lambda(1-2\theta)}{4(1-\theta)^2\lambda}$ | $\frac{2\theta^3(1-\theta)\lambda + (1-2\theta)(1-2\theta+2\theta^2)\lambda^2 - \theta^2(1-\theta)^2}{16(1-\theta)^4\lambda^2}$ |
| **2** | $i$ | $1-i$ | $i$ | $1-i$ | $\frac{2(1-\theta)^2\lambda - \theta(1-\theta) - \lambda(1-2\theta)}{4(1-\theta)^2\lambda}$ | $\frac{2\theta^3(1-\theta)\lambda + (1-2\theta)(1-2\theta+2\theta^2)\lambda^2 - \theta^2(1-\theta)^2}{16(1-\theta)^4\lambda^2}$ |
| **3** | $i$ | $i$ | $1-i$ | $1-i$ | $\frac{2\theta^2\lambda + \theta(1-\theta) - \lambda(1-2\theta)}{4\theta^2\lambda}$ | $\frac{2\theta(1-\theta)^3\lambda - (1-2\theta)(1-2\theta+2\theta^2)\lambda^2 - \theta^2(1-\theta)^2}{16\theta^4\lambda^2}$ |
| **4** | $i$ | $1-i$ | $1-i$ | $I$ | $\frac{2\theta^2\lambda + \theta(1-\theta) + \lambda(1-2\theta)}{4\theta^2\lambda}$ | $\frac{2\theta(1-\theta)^3\lambda - (1-2\theta)(1-2\theta+2\theta^2)\lambda^2 - \theta^2(1-\theta)^2}{16\theta^4\lambda^2}$ |
| **5** | $1-i$<br>$i$<br>$i$<br>$i$ | $i$<br>$1-i$<br>$i$<br>$i$ | $i$<br>$i$<br>$1-i$<br>$i$ | $I$<br>$I$<br>$I$<br>$1-i$ | $1/2$ | $\frac{\theta^2\lambda + (1-\theta)^2\lambda - \theta(1-\theta)}{16\theta(1-\theta)\lambda^2}$ |
| **6** | $i$<br>… | $i$<br>… | …<br>$i$ | …<br>$I$ | $\frac{\lambda + \theta(1-\theta)}{2\lambda}$ | $\frac{\lambda - 2\theta^2(1-\theta)^2 - \theta(1-\theta)}{8\lambda^2}$ |
| **7** | $i$<br>… | $1-i$<br>… | …<br>$i$ | …<br>$1-i$ | $\frac{\lambda - \theta(1-\theta)}{2\lambda}$ | $\frac{\lambda - 2\theta^2(1-\theta)^2 - \theta(1-\theta)}{8\lambda^2}$ |
| **8** | …<br>$i$<br>$i$<br>$i$ | $i$<br>…<br>$i$<br>$i$ | $i$<br>$i$<br>…<br>$i$ | $I$<br>$I$<br>$i$<br>… | $\frac{(3-4\theta)\lambda + \theta(1-\theta)}{4\lambda(1-\theta)}$ | $\frac{(1-\theta)(1-2\theta+4\theta^2)\lambda - \theta(1-\theta)^2(1+\theta) + (1-2\theta)\lambda^2}{16(1-\theta)^2\lambda^2}$ |
| **9** | …<br>$i$<br>$i$<br>$1-i$ | $i$<br>…<br>$1-i$<br>$i$ | $1-i$<br>$i$<br>…<br>$1-i$ | $I$<br>$1-i$<br>$1-i$<br>… | $\frac{\lambda - \theta(1-\theta)}{4\lambda(1-\theta)}$ | $\frac{(1-\theta)(1-2\theta+4\theta^2)\lambda - \theta(1-\theta)^2(1+\theta) + (1-2\theta)\lambda^2}{16(1-\theta)^2\lambda^2}$ |
| **10** | …<br>$i$<br>$i$<br>$i$ | $i$<br>…<br>$i$<br>$i$ | $1-i$<br>$1-i$<br>$1-i$<br>… | $1-i$<br>$1-i$<br>…<br>$1-i$ | $\frac{(4\theta-1)\lambda + \theta(1-\theta)}{4\theta\lambda}$ | $\frac{\theta(3-6\theta+4\theta^2)\lambda - \theta^2(1-\theta)(2-\theta) - (1-2\theta)\lambda^2}{16\theta^2\lambda^2}$ |
| **11** | …<br>$i$<br>$i$<br>$1-i$ | $i$<br>…<br>$1-i$<br>$i$ | $i$<br>$1-i$<br>$1-i$<br>… | $1-i$<br>$I$<br>…<br>$1-i$ | $\frac{\lambda - \theta(1-\theta)}{4\lambda\theta}$ | $\frac{\theta(3-6\theta+4\theta^2)\lambda - \theta^2(1-\theta)(2-\theta) - (1-2\theta)\lambda^2}{16\theta^2\lambda^2}$ |
| **12** | $i$<br>… | …<br>$i$ | …<br>$i$ | $I$<br>… | $1-\theta$ | $\frac{\lambda - \theta(1-\theta) - 2(1-2\theta)^2\lambda^2}{8\lambda^2}$ |
| **13** | $i$<br>… | …<br>$i$ | …<br>$1-i$ | $1-i$<br>… | $\theta$ | $\frac{\lambda - \theta(1-\theta) - 2(1-2\theta)^2\lambda^2}{8\lambda^2}$ |
| **14** | $i$<br>… | …<br>$i$ | $i$<br>… | …<br>$I$ | $1/2$ | $\frac{\lambda - \theta(1-\theta) - 2(1-2\theta)\lambda^2}{16(1-\theta)\lambda^2}$ |
| **15** | $i$<br>… | …<br>$i$ | $1-i$<br>… | …<br>$1-i$ | $1/2$ | $\frac{\lambda - \theta(1-\theta) - 2(1-2\theta)\lambda^2}{16\theta\lambda^2}$ |
| **16** | $i$<br>…<br>…<br>… | …<br>$i$<br>…<br>… | …<br>…<br>$i$<br>… | …<br>…<br>…<br>$I$ | $1/2$ | $\frac{\lambda - \theta(1-\theta)}{8\lambda^2}$ |

**Ejemplo**

En la Figura 2.2 se presenta un ejemplo ilustrativo de cómo funciona el método propuesto por Guo (1994).

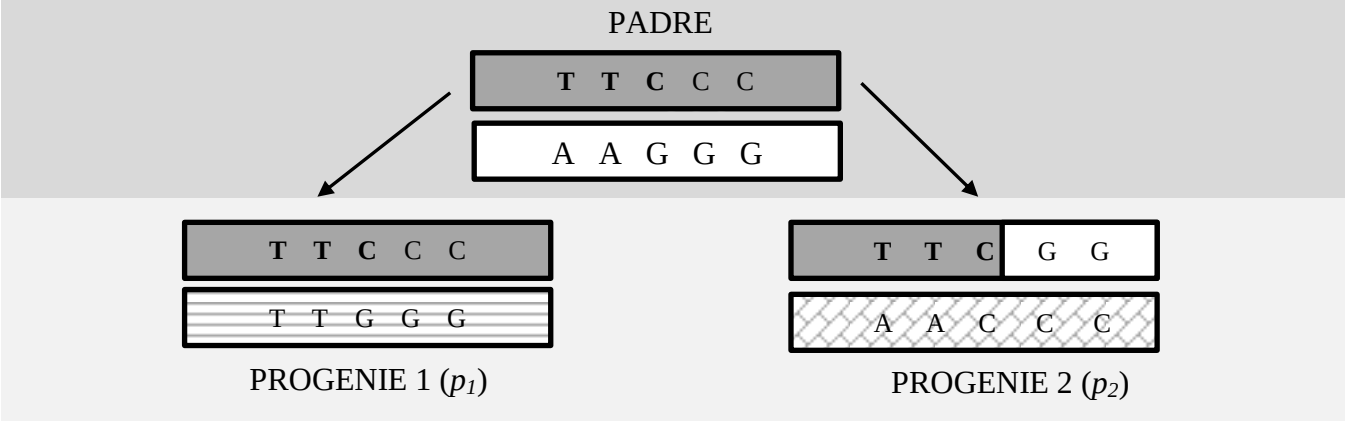



**Figura 2.2.** Ejemplo para una pareja de medio-hermanos paternos.

El ejemplo de la Figura 2.2 ilustra un caso de medio-hermanos paternos, extrapolable al caso de medio-hermanos maternos. Las barras con diferente sombreado representan parte de un cromosoma con 5 marcadores SNPs cada uno. Dichos marcadores definen 4 segmentos de 0,05 Morgans cada uno (definidos entre el marcador 1 y 2, entre el 2 y 3, entre 3 y 4 y entre 4 y 5). Cada uno de los hijos ($p_1$ y $p_2$) recibió diferentes segmentos del cromosoma paterno, tal como lo representa la barra superior en cada uno de ellos. La barra inferior corresponde al segmento cromosómico heredado de la madre de cada individuo, que en este caso son diferentes en ambos animales. Las mismas se asumen no emparentas y, en consecuencia, no existe la posibilidad que $p_1$ y $p_2$ compartan genoma IBD en aquellos segmentos heredados de sus madres. Ahora bien, para aquellas regiones cromosómicas provenientes del padre es necesario determinar cuáles son los casos presentados en el Cuadro 2.1 que aplican en cada segmento. A tal efecto es necesario considerar el genoma del padre conjuntamente con aquellos segmentos heredados por sus hijos por vía paterna, y representados en la figura por las barras superiores del genoma de $p_1$ y $p_2$ (la fase). Ahora bien, nótese que para el primer segmento definido entre los marcadores 1 y 2, ambos individuos recibieron los mismos alelos paternos: T y T. Es decir que $i_1 = i_2 = j_1 = j_2 = 1$ correspondiendo al caso 1 del Cuadro 2.1, motivo por el cual la probabilidad de que ambos segmentos sean IBD se calcula del siguiente modo $PIBD_{p_1,p_2(1)} = \frac{2(1-\theta)^2\lambda + \theta(1-\theta) + \lambda(1-2\theta)}{4(1-\theta)^2\lambda} = 0,99924$. Para el segundo segmento definido por los marcadores 2 y 3, ocurre la misma situación que con el segmento precedente, y se observan que ambos individuos heredaron los mismos alelos por vía paterna (T y C). Nuevamente se cumple que $i_1 = i_2 = j_1 = j_2 = 1$ y la probabilidad de que ambos segmentos sean IBD es

$PIBD_{p_1,p_2(2)} = \frac{2(1-\theta)^2\lambda + \theta(1-\theta) + \lambda(1-2\theta)}{4(1-\theta)^2\lambda} = 0,99924$. En el caso del tercer segmento entre los marcadores 3 y 4 la situación es diferente. Ocurre que ambos individuos recibieron el mismo alelo del padre en el marcador 3 (C), pero el marcador 4 fue distinto en ambos individuos: $p_1$ recibió el alelo proveniente del abuelo paterno (C), mientras que $p_2$ recibió el alelo proveniente de la abuela paterna (G). Entonces, $i_1 = i_2 = 1$ pero $j_1 \neq j_2$, $j_1 = 1$ y $j_2 = 0$ situación que corresponde al caso 5 del Cuadro 2.1. La probabilidad de que ambos segmentos sean IBD es entonces igual a $PIBD_{p_1,p_2(3)} = 0,5$ tal como se observa en el Cuadro 2.1. Finalmente, para el último segmento definido entre los marcadores 4 y 5, ambos individuos recibieron alelos diferentes del padre: el individuo $p_1$ recibió alelos provenientes del abuelo paterno para ambos marcadores (C, C) y $p_2$ recibió los alelos provenientes de la abuela paterna (G, G). En este caso $i_1 \neq i_2$ y $j_1 \neq j_2$, donde $i_1 = j_1 = 1$ y $i_2 = j_2 = 0$, con probabilidad de que ambos segmentos sean IBD igual a $PIBD_{p_1,p_2(4)} = \frac{2(1-\theta)^2\lambda - \theta(1-\theta) - \lambda(1-2\theta)}{4(1-\theta)^2\lambda} = 0,000827$. Una vez calculadas las probabilidades de IBD para cada segmento se obtiene un valor global de PIBD, que en el ejemplo corresponde al segmento cromosómico completo con los 5 marcadores, del modo siguiente

$$\text{PIBD}_{p_1,p_2\,\text{global}} = \frac{\sum_{i=1}^{N}\text{PIBD}_i}{N} = \frac{0,99924 + 0,99924 + 0,5 + 0,000827 + 0 + 0 + 0 + 0}{8} = 0,3124 \quad [2]$$

donde *N* es el número total de segmentos considerados. Esta forma de cálculo de la PIBD global aplica solamente para este caso excepcional en el que todos los segmentos tienen la misma longitud. Para aquellos casos en los que los segmentos tienen longitudes variables será necesario ponderar cada PIBD por el largo del segmento y dividir la suma ponderada resultante por el largo total del genoma, tal como se indica en la expresión [1]. Nótese que en el caso del ejemplo descripto, la división es por 8 a la hora de calcular la PIBD global dado que para determinar qué proporción del segmento cromosómico comparten, no sólo es necesario considerar los cuatro segmentos de origen paterno, sino también aquellos otros cuatro provenientes de las madres, donde en todos los casos la probabilidad de IBD es igual a 0. De la estimación de la PIBD global se infiere entonces que ambos medio-hermanos comparten aproximadamente un 31% de su genoma en la región cromosómica analizada.

### 2.3. El método de Han y Abney (2011) y los modelos ocultos de Markov

Han y Abney (2011) propusieron un método alternativo para estimar las probabilidades de IBD en pares de individuos emparentados condicionando en la información de marcadores moleculares y de la genealogía. Si bien existen métodos similares, tales como los propuestos por Albrechtsen *et al.* (2009) o por Purcell *et al.* (2007), la ventaja del método presentado por Han y Abney (2011) es que considera el LD

entre marcadores y la consanguinidad. Dicha metodología emplea un HMM, motivo por el cual en esta sección se realizará, en una primera instancia, una breve descripción de las características de estos modelos estadísticos para posteriormente abordar con mayor grado de detalle el modo en que Han y Abney (2011) propusieron estimar las probabilidades de IBD.

### 2.3.1. Modelos ocultos de Markov

Un HMM permite modelar un proceso estocástico que genera una sucesión de observaciones. Consta de: 1) un grupo de variables aleatorias no observadas que siguen un proceso de Markov; en nuestro caso:

$$S = \left\{ \left( X \equiv Y \right)_1, \left( X \equiv Y \right)_2, ..., \left( X \equiv Y \right)_l \right\} = \left\{ S_1, S_2, ..., S_l \right\} \qquad [3]$$

Las variables $\left( X \equiv Y \right)_i$ son los eventos que surgen al observar IBD entre los individuos *X* e *Y* en el marcador *i*, donde *i* toma valores de 1 a *l* (número total de marcadores); y 2) un grupo de variables aleatorias observadas que resultan de la secuencia *S*, y que llamaremos $M = \left\{ M_1, M_2, ..., M_n \right\}$. En nuestro caso $l = n$, dado que se cuenta con una observación (genotipo observado) para cada marcador.

Un proceso estocástico es considerado Markoviano cuando la distribución de la variable $S_i$ en el estado *i* depende solamente de la variable inmediatamente anterior, $S_{i-1}$. Dicho de otro modo, sea $s_i$ el estado IBD realizado en la posición *i*, la probabilidad de transición hacia *i* + 1 es igual a $P\left( S_{i+1} = s_{i+1} \middle| S_i = s_i \right)$ y, por la propiedad antes mencionada, se cumple

$$P\left( S_{i+1} = s_{i+1} \middle| S_i = s_i, S_{i-1} = s_{i-1}, ..., S_{i-l} = s_{i-l} \right) = P\left( S_{i+1} = s_{i+1} \middle| S_i = s_i \right) \qquad [4]$$

En otras palabras, existe independencia condicional para la sucesión de estados porque, dado $s_i$, la variable $s_{i+1}$ no depende de ningún otro estado previo, y es consecuentemente llamada cadena de Markov de "primer orden". Formalmente, un modelo oculto de Markov se define sobre la base de cinco elementos:

1. Un conjunto de *l* estados ocultos, que para el método propuesto por Han y Abney (2011) corresponden a los modos IBD presentados por Jacquard (1974, pág. 104), tal que $l = 9,$ con lo cual: $s_i = \left\{ 1, 2, ..., 9 \right\}$.
2. Un conjunto de *r* símbolos observables para cada estado, definido de la siguiente manera: $M_i = \left\{ m_1, m_2, ..., m_r \right\}$. En nuestro caso los símbolos corresponden a los alelos presentes en cada SNP. Es importante destacar que cada marcador genera una sola

observación, es decir que para cada SNP se puede observar A, T, C o G: $M_i = \{m_1 = \text{A}, m_2 = T, m_3 = C, m_4 = G\}$.

3. Un vector $\boldsymbol{\pi}$ de orden $l \times 1$, de probabilidades *iniciales* en el primer marcador ($S_1$), tal que el elemento *i* de $\boldsymbol{\pi}$ es igual a $\pi_i = P(S_1 = s_i)$, *i* = 1,..., *l*.
4. La matriz ***T*** de probabilidades de transición entre estados, de orden $l \times l$ cuyos elementos son iguales a $T_{jk} = P(S_{i+1} = s_k \mid S_i = s_j)$ siendo $j = 1,\ldots, l$ y $k = 1,\ldots, l$.
5. La matriz ***B*** de probabilidades de *emisión* de los genotipos observables, de orden $l \times r$. Sus elementos son $B_j(k) = P(\text{genotipo } m_k \text{ en la posición } i \mid S_i = s_j)$ donde $b_j(k)$ representa la probabilidad de emisión del genotipo *k* en el estado *j*, donde $j = 1,\ldots, l$ y $k = 1,\ldots, r$.

Los HMM son modelos "generativos" porque producen o "emiten" secuencias de variables observables. El proceso comienza en un estado determinado al que se le asocia una probabilidad de iniciación y genera una observación, para posteriormente transitar hacia un nuevo estado en el marcador siguiente, el cual "emite" una nueva observación. Los sistemas HMM evolucionan a lo largo del genoma, transitando entre estados y emitiendo en cada posición una observación (un valor de $M_i$), hasta alcanzar el final de la secuencia. Se los conoce con el calificativo de "ocultos" dado que la secuencia de estados por la cual transita el sistema no es observable, pero sí lo son las observaciones emitidas las cuales son independientes entre sí. Dada $s = \{s_1, s_2, \ldots, s_t\}$, la probabilidad de observar una secuencia determinada de *t* observaciones, $m = \{m_1, m_2, \ldots, m_t\}$ es igual a:

$$P(m \mid s) = \prod_{i=1}^{t} P(m_t \mid s_i) \qquad [5]$$

Desarrollando la expresión [5] se obtiene

$$P(m \mid s) = B_{s_1}(m_1) \times B_{s_2}(m_2) \times \ \ldots \times B_{s_t}(m_t) \qquad [6]$$

La probabilidad de la secuencia de estados (*s*) se calcula de la manera siguiente:

$$P(s) = \pi_{s_1} \times T_{s_1 s_2} \times T_{s_2 s_3} \times T_{s_3 s_4} \times \ \ldots \times T_{s_{t-1} s_t} \qquad [7]$$

Por lo que la probabilidad conjunta de *m* y *s*, es decir la probabilidad que la secuencia *m* y *s* ocurran simultáneamente, se obtiene multiplicando las expresiones [6] y [7]:

$$P(m, s) = P(m \mid s) \times P(s) \qquad [8]$$

Ahora bien, una de las probabilidades de interés es la de observar la secuencia *m*. Dado que es una probabilidad marginal, se calcula sumando la distribución conjunta sobre todos las posibles secuencias de estados *s*, del siguiente modo:

$$
\begin{aligned}
P(m) &= \sum_{s_1, s_2, \ldots, s_t} P(m,s) \\
&= \sum_{s_1, s_2, \ldots, s_t} P(m \mid s) \times P(s) \\
&= \sum_{q_1, q_2, \ldots, q_t} \pi_{s_1} B_{s_1}(m_t) \prod_{i=2}^{t} B_{s_i}(m_i) T_{s_{i-1} s_i}
\end{aligned}
\qquad [9]
$$

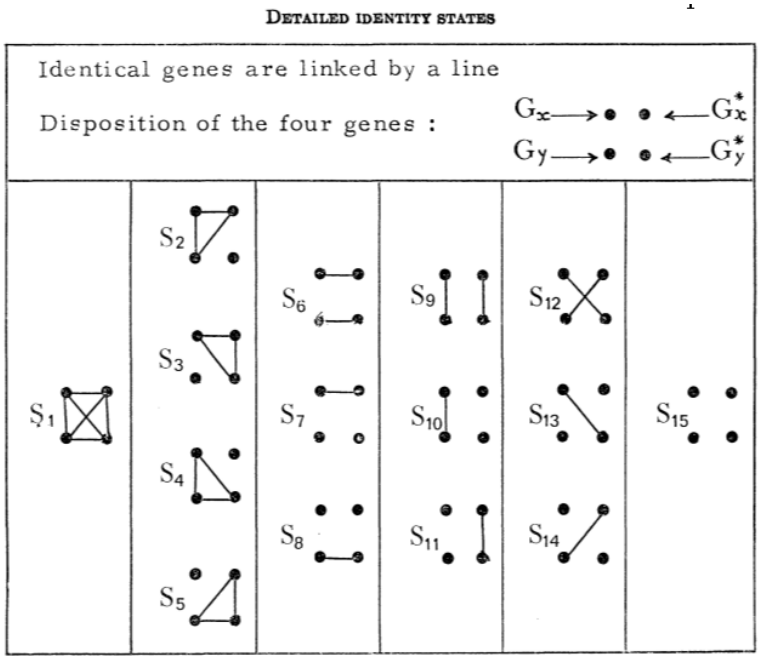
La expresión [9] pone a consideración la manera en que evoluciona el sistema desde el inicio de la secuencia hasta alcanzar el final. Inicialmente, en el primer marcador, el proceso se encuentra en el estado $s_1$ cuya probabilidad es $\pi_{s_1}$, a partir del cual se genera la observación $m_1$ con probabilidad $B_{s_1}(m_1)$. Al avanzar a la posición $i = 2$ se produce una transición del estado $s_1$ al $s_2$, a la cual se asocia $T_{s_1 s_2}$, obteniéndose como resultado el genotipo $m_2$ con probabilidad $B_{s_2}(m_2)$. El proceso continúa de este modo hasta llegar al final de la secuencia, que se alcanza en el estado $s_t$.

Los HMM permiten obtener estimaciones multipunto de la probabilidad de IBD para cada par de individuos, dado el patrón de identidad en estado observado. Si bien los estados de IBD a lo largo del cromosoma no siguen exactamente un proceso markoviano, McPeek y Sun (2000) mostraron que los HMM producen una buena aproximación a la PIBD. Los HMM estándar utilizados para calcular dicha proporción asumen ausencia de consanguinidad; en otras palabras, que los alelos dentro de un mismo individuo no son IBD. Tal es el caso de la metodología propuesta por Purcell *et al.* (2007) y Albrechtesen *et al.* (2009), donde sólo se consideran tres estados de identidad posibles: cero, uno o dos alelos idénticos por descendencia compartidos por el par de individuos. Estos dos algoritmos permiten calcular la probabilidad condicional de IBD en el espacio $s = 0, 1, 2$, dados los genotipos $M$ de los $k$ marcadores en un cromosoma, formalmente $P(S = s \mid M)$.

### 2.3.2. El modelo de Han y Abney (2011)

Han y Abney (2011) presentaron un enfoque alternativo modificando el algoritmo de Purcell *et al.* (2007), para incorporar la consanguinidad y el LD en los cálculos. Para ello definieron la variable de estado oculto $S_i$ asociada con el marcador $i$, que puede tomar uno entre nueve posibles valores, en función del estado de identidad "condensado" de Jacquard (1974, p. 105). Cada estado de identidad condensado corresponde a uno o más de los 15 modos de identidad presentados en el Cuadro 2.2. Por ejemplo $S_8'$ corresponde a $S_{10}, S_{11}, S_{13}$ y $S_{14}$. Los nueve casos posibles se presentan en el Cuadro 2.3.

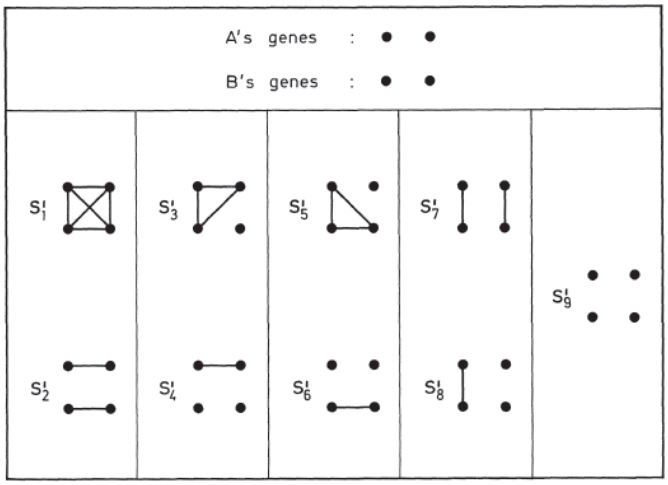
**Cuadro 2.2.** Modos de identidad presentados por Jacquard (1974, p. 105). En cada grupo de cuatro puntos, aquellos dos superiores representan los genes del individuo A, mientras que los inferiores son del individuo B. Los genes idénticos se encuentran unidos por una línea.




En ausencia de consanguinidad, los estados $S_1$ a $S_8$, en el cuadro 2.2, son imposibles. Entonces

$$\omega_0 = P(S_{15}) \qquad \omega_1 = P(S_{10}) + P(S_{11}) + P(S_{13}) + P(S_{14}) \qquad \omega_2 = P(S_9) + P(S_{12}) \qquad [10]$$

**Cuadro 2.3.** Modos de identidad condensados presentados por Jacquard (1974, p. 107). En cada grupo de cuatro puntos, los dos superiores representan los genes del individuo A y los inferiores, los del individuo B. Los genes idénticos se encuentran unidos por una línea.



Cuando el pedigrí es conocido, se le asigna una probabilidad $(\Delta_i)$ a cada estado condensado, y se obtienen los siguientes "coeficientes de identidad por descendencia condensados" (Jacquard, 1974, p. 106):

$$\Delta_1 = \delta_1 \qquad \Delta_6 = \delta_8$$

$$\Delta_2 = \delta_6 \qquad \Delta_7 = \delta_9 + \delta_{12}$$

$$\Delta_3 = \delta_2 + \delta_3 \qquad \Delta_8 = \delta_{10} + \delta_{11} + \delta_{13} + \delta_{14}$$

$$\Delta_4 = \delta_7 \qquad \Delta_9 = \delta_{15}$$

$$\Delta_5 = \delta_4 + \delta_5$$

El valor $\delta_i$ corresponde, en cada caso, a la probabilidad del modo de identidad *i*, donde *i* toma valores de 1 a 15.

Consecuentemente, la probabilidad *a priori* en el primer marcador está determinada por el coeficiente de IBD condensado asociado con el modo puntual observado $(\Delta_i)$, *i* = 1, …, 9. Se considera que la secuencia de estados de IBD para los *L* marcadores $(S_1,...,S_L)$ sigue un modelo estocástico de cadenas de Markov en el que la distribución de los estados de identidad en el marcador *i* + 1 depende solamente de la matriz de transición:

$$\boldsymbol{T}_{rt} = P\left(S_{i+1} = t \mid S_i = r\right) \qquad [11]$$

El estado de identidad del marcador *i* se denota como *r*, siendo *t* el estado de identidad en el marcador *i* + 1. El último elemento de la cadena de Markov está dado por las probabilidades de los genotipos del par de individuos, las cuales son condicionales al valor del estado de identidad subyacente (probabilidades de emisión): $P\left(M_i \mid S_i = s_i\right)$. Si las tres probabilidades son conocidas (la *a priori*, $\boldsymbol{T}_{rt} = P\left(S_{i+1} = t \mid S_i = r\right)$ y $P\left(M_i \mid S_i = s_i\right)$) es posible estimar las probabilidades de cada estado de identidad en un punto cualquiera del cromosoma dada la información de los genotipos del par considerado: $P\left(S_i \mid M_i\right)$. Para calcular $P\left(M_i \mid S_i = s_i\right)$ se utiliza el algoritmo "forward-backward" (Rabiner, 1989; Durbin *et al.*, 1998).

Ahora bien, las probabilidades de transición ($\boldsymbol{T}_{rt} = P\left(S_{i+1} = t \mid S_i = r\right)$) dependen de la distancia genética entre los marcadores, así también como del pedigrí del par de individuos considerados. Han y Abney (2011) propusieron estimarlas definiendo un vector $\boldsymbol{s}_i = \left(1_{S_i=1},...,1_{S_i=9}\right)$ cuyos elementos son funciones indicadoras del estado de IBD en la posición *i*. Así, la distribución de probabilidad en el marcador *i* + 1 es igual a:

$$P\left(\boldsymbol{s}_{i+1}\right) = P\left(\boldsymbol{s}_i\right)\boldsymbol{T}\left(x_{i+1} - x_i\right) \qquad [12]$$

La matriz de probabilidades de transición para la distancia genética *x* es $\boldsymbol{T}\left(x\right)$. Debemos resaltar dos aspectos en la expresión [12]: 1) la dependencia existente entre el estado actual y el inmediatamente anterior; 2) la intervención de la distancia genética entre los marcadores en las probabilidades de transición. Al asumir que los estados ocultos forman una cadena de Markov, la matriz de transición puede expresarse como

$$\boldsymbol{T}\left(x\right) = \mathrm{e}^{\boldsymbol{Q}x} \qquad [13]$$

donde ***Q*** es la matriz "infinitesimal" (Norris, 1997, p. 64). Puede demostrarse que

$$\boldsymbol{T}\left(x\right) = \boldsymbol{U}\ \boldsymbol{D}\left(x\right)\boldsymbol{U}^{-1} \qquad [14]$$

Las columnas de la matriz ***U*** son los autovectores de ***Q***, mientras que ***D***(*x*) es una matriz diagonal con los autovalores de ***Q***. Por este motivo, y teniendo en cuenta la descomposición espectral de la matriz de transición, puede demostrarse que la probabilidad de pasar del estado *r* al *t*, dada la distancia *x* entre ambos marcadores, puede expresarse del siguiente modo:

$$\boldsymbol{T}_{rt}\left(x\right) = \sum_{j=1}^{9} A_{j,rt}\ e^{\lambda_j x} \qquad [15]$$

donde los elementos de la matriz de transición están sujetos a las siguientes condiciones

$$\boldsymbol{T}_{\boldsymbol{rt}}\left(0\right) = \begin{cases} 0 & r \neq t \\ 1 & r = t \end{cases} \quad \text{y} \quad \lim_{x\to\infty} \boldsymbol{T}_{\boldsymbol{rt}}\left(x\right) = \Delta_t \qquad [16]$$

Una manera de aproximar las probabilidades de transición es la siguiente:

$$\boldsymbol{T}_{\boldsymbol{ij}}\left(x\right) \approx \boldsymbol{A}_{1,ij} + \boldsymbol{A}_{2,ij}\, e^{\lambda x} \qquad [17]$$

donde $a_{1,\mathrm{ij}} = \Delta_\mathrm{j}$ y $a_{2,\mathrm{ij}} = -\Delta_\mathrm{j}$ para $i \neq j$ o $a_{1,\mathrm{ii}} = \Delta_\mathrm{i}$ y $a_{2,\mathrm{ii}} = 1-\Delta_\mathrm{i}$ para $i = j$.

Las probabilidades de emisión son aquellas relacionadas con los genotipos observados para el par de individuos considerado dado su estado de identidad condensado en el locus *i*, $P\left(M_i \mid S_i = s_i\right)$. En este caso se asumen que los SNPs son bialélicos, codificados con 0 y 1. El genotipo del individuo *p* se denota como $M_i^p$ y es igual a 0, 1 o 2 alelos de referencia. Las probabilidades de emisión dependen de las frecuencias alélicas del locus considerado; están calculadas para cada estado de identidad condensado (Han y Abney, 2011) y se presentan en el Cuadro 2.4. Se asume que las probabilidades genotípicas para ambos individuos (A y B) son idénticas, e iguales al producto de las frecuencias

alélicas. Si estas condiciones no se cumplen es necesario considerar cada caso a nivel de genotipo en cada individuo.

**Cuadro 2.4.** Probabilidades de los genotipos dado los estados condensados de identidad. Tomado de Han y Abney (2011). $f_i$ es la frecuencia alélica del alelo *i* (*i* = *a, b, c, d*).

| $S$ | $Pr(G^1 = (a,b), G^2 = (c,d)\|S)^\dagger$ |
|---|---|
| 1 | $\delta_{ab}\delta_{ac}\delta_{ad}f_a$ |
| 2 | $\delta_{ab}\delta_{cd}f_af_c$ |
| 3 | $\frac{1}{2}(2-\delta_{cd})\delta_{ab}(\delta_{ac}f_af_d+\delta_{ad}f_af_c)$ |
| 4 | $(2-\delta_{cd})\delta_{ab}f_af_cf_d$ |
| 5 | $\frac{1}{2}(2-\delta_{ab})\delta_{cd}(\delta_{ca}f_af_b+\delta_{cb}f_cf_a)$ |
| 6 | $(2-\delta_{ab})\delta_{cd}f_cf_af_b$ |
| 7 | $\frac{1}{2}(2-\delta_{ab})(2-\delta_{cd})(\delta_{ac}\delta_{bd}f_af_b+\delta_{ad}\delta_{bc}f_af_b)$ |
| 8 | $\frac{1}{4}(2-\delta_{ab})(2-\delta_{cd})(\delta_{ac}f_af_bf_d+\delta_{ad}f_af_bf_c+\delta_{bc}f_bf_af_d+\delta_{bd}f_bf_af_c)$ |
| 9 | $(2-\delta_{ab})(2-\delta_{cd})f_af_bf_cf_d$ |

Nótese que el delta de Kronecker ($\delta_{ii'}$) involucra la noción de identidad en estado. Permite comparar cada uno de los alelos de los marcadores entre ellos, dentro y entre individuos, y al detectar IBS asigna un valor 1 (o cero en cualquier otro caso). Cabe destacar, además, que los deltas de Kronecker utilizados dependen del estado de identidad considerado. Por ejemplo, dada las características del estado de identidad 1, es necesario utilizar aquellas funciones que contemplen la identidad entre todos los alelos. En este caso, y dado que todos los pares de genes son IBD, su frecuencia es igual a $f_a$. La probabilidad de emisión es entonces el producto de los deltas de Kronecker que relacionan a todos los alelos entre sí y la frecuencia alélica. La misma lógica se aplica para los ocho estados de identidad restantes, en los que se consideran los alelos de interés según cada caso y las frecuencias alélicas correspondientes.

Ahora bien, hasta aquí se ha trabajado sobre el supuesto de que el genotipo observado es el verdadero del individuo. Pero en muchas ocasiones esto no se cumple debido a errores de genotipado o a información faltante producto de fallas de diversa índole. Conocer esta situación permite tenerla en cuenta a la hora de trabajar con información genómica. Para ello se incluye un conjunto adicional de probabilidades que permiten modelar dichos efectos. Sea $O_i = \left(O_i^1, O_i^2\right)$ el genotipo observado en el marcador *i* para los dos individuos y $O_{\text{M}} = \left(O_{\text{M}}^1, O_{\text{M}}^2\right)$ el conjunto de genotipos faltantes para el par de animales. Las probabilidades ($O_i^P$) de observar un cierto genotipo en el marcador *i*, son condicionales a los genotipos faltantes ($O_{\text{M}}$), y se asume que el mecanismo de pérdida de información (valores faltantes) es independiente del genotipo subyacente: $P\left(O_i^P = - \middle| O_{\text{M}}\right) = 1$ (el símbolo "−" representa un genotipo faltante). El error de genotipado es considerado mediante la introducción del parámetro ε y se utilizan las probabilidades propuestas por Han y Abney (2011), tal como se presentan en el Cuadro 2.5.

**Cuadro 2.5.** Probabilidades de los genotipos observados dado los reales, considerando el error de genotipado (ε). Tomado de Han y Abney (2011).

| $P(O_i^p \mid G_i^p, \epsilon)$ | $O_i^p = 0$ | $O_i^p = 1$ | $O_i^p = 2$ |
|---|---|---|---|
| $G_i^p = 0$ | $(1-\epsilon)^2$ | $2\epsilon(1-\epsilon)$ | $\epsilon^2$ |
| $G_i^p = 1$ | $\epsilon(1-\epsilon)$ | $\epsilon^2 + (1-\epsilon)^2$ | $\epsilon(1-\epsilon)$ |
| $G_i^p = 2$ | $\epsilon^2$ | $2\epsilon(1-\epsilon)$ | $(1-\epsilon)^2$ |

### 2.3.2.1. Modelo para el desequilibrio de ligamiento

Existen diversas alternativas a considerar para modelar el LD. Se lo suele incluir en un HMM a partir de las probabilidades de emisión y existen diversos modos de hacerlo. Se destacan dos modos contrastantes: 1) condicionar la probabilidad de los genotipos en el de un sólo SNP, o 2) condicionarla en varios SNPs. El primer enfoque es aplicable a casos en que el LD entre dos SNPs es alto pero no se encuentran asociados a otros marcadores. Estos casos suelen presentarse con muy baja frecuencia. En general, existe escaso LD entre pares de marcadores y una alta dependencia entre varios loci, motivo por el cual resulta más adecuado utilizar el segundo enfoque. Ahora bien, con el objetivo de utilizar un método computacionalmente eficiente, se recurre a un modelo lineal propuesto por Han y Abney (2011), que permite aproximar la estructura del LD y corregir el HMM por dicho efecto a lo largo de *L* loci. A modo de ejemplo se puede evaluar la probabilidad condicional del individuo 1, cuyo genotipo en el marcador *i* es 0, dado el genotipo del marcador anterior (*i* – 1):

$$P\left(\mathrm{M}_i^1 = 0 \middle| \mathrm{M}_{i-1}^1\right) = P\left(\mathrm{M}_i^1 = 0 \middle| \mathrm{M}_{i-1}^1 = 0\right)1_{\mathrm{M}_{i-1}^1=0} + P\left(\mathrm{M}_i^1 = 0 \middle| \mathrm{M}_{i-1}^1 = 1\right)1_{\mathrm{M}_{i-1}^1=1} + \quad [18]$$

$$P\left(\mathrm{M}_i^1 = 0 \middle| \mathrm{M}_{i-1}^1 = 2\right)1_{\mathrm{M}_{i-1}^1=2} \quad = \gamma_{i,0} + \gamma_{i-1,00}1_{\mathrm{M}_{i-1}^1=0} + \gamma_{i-1,20}1_{\mathrm{M}_{i-1}^1=2}$$

Las funciones indicadoras $\left(1_{\mathrm{M}_{i-1}^1=i}\right)$ son iguales a 1 cada vez que se cumpla la condición $\mathrm{M}_{i-1}^1 = i$ o iguales a 0 en caso contrario. Los parámetros quedan definidos del siguiente modo:

$$\gamma_{i,0} = P\left(\mathrm{M}_i^1 = 0 \middle| \mathrm{M}_{i-1}^1 = 1\right), \gamma_{i-1,00} = P\left(\mathrm{M}_i^1 = 0 \middle| \mathrm{M}_{i-1}^1 = 0\right) - P\left(\mathrm{M}_i^1 = 0 \middle| \mathrm{M}_{i-1}^1 = 1\right)$$
$$\gamma_{i-1,20} = P\left(\mathrm{M}_i^1 = 0 \middle| \mathrm{M}_{i-1}^1 = 2\right) - P\left(\mathrm{M}_i^1 = 0 \middle| \mathrm{M}_{i-1}^1 = 1\right) \quad [19]$$

Ahora bien, para extender este modelo a *L* loci, se define el genotipo del individuo *p* en el locus *i* en términos de la función indicadora que considera los dos genotipos homocigotas,

$\mathrm{M}_i^p = \left(1_{\mathrm{M}_i^p=0}, 1_{\mathrm{M}_i^p=2}\right)'$. Consecuentemente, la probabilidad condicional del genotipo en el marcador *i* se calcula a partir de la siguiente expresión (Han y Abney, 2011):

$$P\left(\mathrm{M}_i^p \middle| \mathrm{M}_{i-1}^p,...,\mathrm{M}_{i-L}^p\right) = \begin{pmatrix} P\left(M_i^p = 0 \middle| \mathrm{M}_{i-1}^p,...,\mathrm{M}_{i-L}^p\right) \\ P\left(M_i^p = 2 \middle| \mathrm{M}_{i-1}^p,...,\mathrm{M}_{i-L}^p\right) \end{pmatrix} \quad [20]$$
$$= \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}\begin{pmatrix} \gamma_{i,0} \\ \gamma_{i,2} \end{pmatrix} + \begin{pmatrix} \gamma_{i-1,00} & \gamma_{i,20} \\ \gamma_{i,02} & \gamma_{i,22} \end{pmatrix}\mathrm{M}_{i-1}^p + ... + \begin{pmatrix} \gamma_{i-L,00} & \gamma_{i-L,20} \\ \gamma_{i-L,02} & \gamma_{i-L,22} \end{pmatrix}\mathrm{M}_{i-L}^p$$

Nótese que se condiciona en un grupo de marcadores adyacentes. Una vez calculadas las probabilidades de los genotipos homocigotas, es posible obtener, la probabilidad del heterocigota como sigue:

$$P\left(M_i^p = 1 \middle| \mathrm{M}_{i-1}^p,...,\mathrm{M}_{i-L}^p\right) = 1 - P\left(M_i^p = 0 \middle| \mathrm{M}_{i-1}^p,...,\mathrm{M}_{i-L}^p\right) - P\left(M_i^p = 2 \middle| \mathrm{M}_{i-1}^p,...,\mathrm{M}_{i-L}^p\right) \quad [21]$$

Estas expresiones permiten calcular las frecuencias de los genotipos en cada locus, condicional a la información de los loci previos, para un individuo. Cabe destacar que las probabilidades de emisión requieren de las probabilidades conjuntas de dos individuos para un locus, dado el estado de identidad subyacente. Para utilizar las probabilidades presentadas en el Cuadro 2.4, será necesario, entonces, convertir las probabilidades genotípicas obtenidas en [20] para cada individuo, a las alélicas específicas.

### 2.3.2.2. Descripción e implementación del algoritmo

El objetivo del algoritmo propuesto es obtener las probabilidades de los estados de identidad dado el genotipo observado $\left(P\left(S = s \middle| \mathrm{O}\right),\ s = 1,\ \ldots,\ 9\right)$. El hecho de contar con estas probabilidades condicionales permite utilizar la metodología propuesta por Jacquard (1974, p. 108) para calcular el coeficiente de coancestría. En el trabajo original (Jacquard, 1974, p. 108) se utilizan las probabilidades de estados de identidad que no están condicionadas a la información de los genotipos. Aquí el cálculo se realiza en función de la información observada del genoma (Han y Abney, 2011). Consecuentemente, se obtiene la PIBD compartida entre ambos individuos. Se lo puede llamar "*coeficiente de coancestría genómico estimado*" para los individuos A y B y se calcula a partir de la siguiente expresión:

$$\hat{\phi}_{\mathrm{AB}} = \frac{1}{K}\sum_{i=1}^{K}\hat{\pi}_i\,, \qquad 0 < \hat{\phi}_{\mathrm{AB}} < 1 \quad [22]$$

*K* representa el número de loci considerados y

$$\hat{\pi}_i = P\left(S=1 \mid \mathrm{O}\right) + \tfrac{1}{2}\left[P\left(S=3 \mid \mathrm{O}\right) + P\left(S=5 \mid \mathrm{O}\right) + P\left(S=7 \mid \mathrm{O}\right)\right] + \tfrac{1}{4}P\left(S=8 \mid \mathrm{O}\right) \qquad [23]$$

El valor $\hat{\pi}_i$ corresponde a la proporción estimada de alelos IBD compartidos por el par de individuos analizados en el locus *i*. La expresión [23] surge al reemplazar las probabilidades marginales de los estados de IBD propuestas originalmente por Jacquard (1974, p. 108) por aquellas condicionales a la información de los genotipos observados.

## 2.4. El método de Elsen *et al.* (2009)

Elsen *et al.* (2009) propusieron un algoritmo eficiente para estimar las probabilidades de transmisión de segmentos cromosómicos de padres a hijos en cada ubicación del genoma. Dichas estimaciones son condicionales a la información de marcadores moleculares SNPs y a las fases parentales. Recordemos que se entiende por fase parental a aquella combinación de alelos que un individuo recibió de sus padres Si bien el método propuesto por Elsen *et al.* (2009) fue desarrollado en el marco de la detección de QTLs, la misma metodología puede ser utilizada para el cálculo de las GWR entre parejas de individuos emparentados. A tal fin se utilizan las probabilidades de transmisión de padres a hijos para calcular, en una primera instancia, la probabilidad de que ambos individuos emparentados hayan heredado el mismo segmento cromosómico para cada ubicación del genoma. Una vez obtenidas dichas probabilidades para cada ubicación del genoma, se procede a calcular una PIBD global teniendo en cuenta cada uno de esos valores, tal como será desarrollado con más detalle más adelante.

La metodología de Elsen *et al.* (2009) contempla cuatro eventos posibles para cada locus autosómico que pueden darse al momento del traspaso de material genético de una generación a la siguiente. Si bien está ampliamente demostrado que cada individuo posee la mitad de su material genético proveniente de su padre y la otra mitad de su madre, no necesariamente cada individuo recibe exactamente un cuarto del genoma de cada uno de sus cuatro abuelos, tal como se ilustra en el ejemplo de la Figura 2.3. Dicho de otro modo, la mitad del genoma paterno que hereda un individuo puede estar compuesto por diferentes proporciones de genoma proveniente de la abuela paterna y del abuelo paterno, al igual que sucede por el lado de la madre, donde el material genético que pasa a su progenie también puede estar formado por diferentes proporciones de genoma que ella recibió de sus padres.

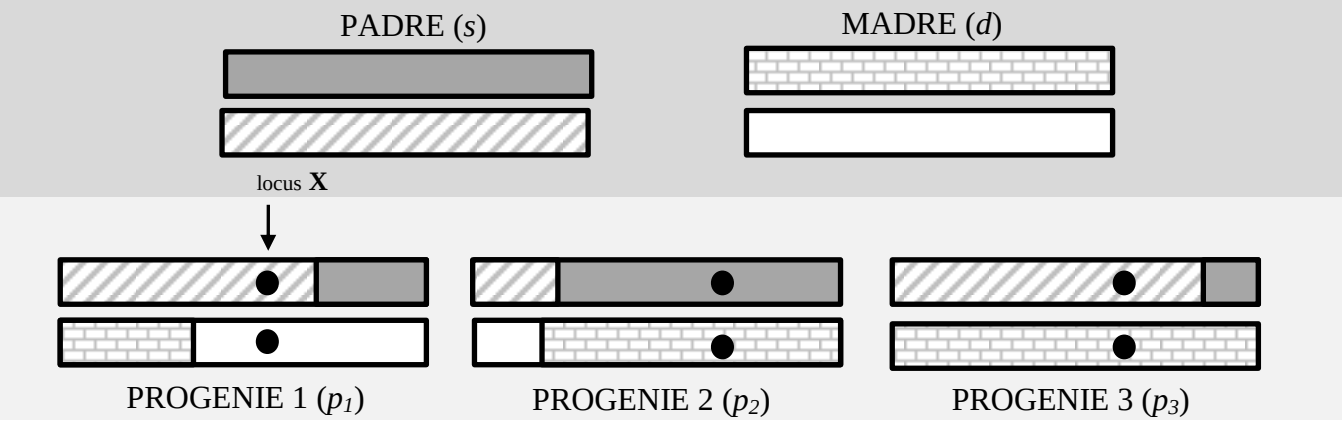



**Figura 2.3.** Ejemplos de posibles eventos de transmisión que pueden ocurrir entre padres e hijos. Cada hijo recibe segmentos cromosómicos de diferente origen para el locus X (provenientes de su abuelo paterno, de su abuela paterna, de su abuelo materno o de su abuela materna).

Tal como se desprende de la Figura 2.3, un individuo puede recibir para cierto locus autosómico (por ejemplo, el locus X) el alelo proveniente de su abuelo paterno, de su abuela paterna, de su abuelo materno o de su abuela materna. Es decir que pueden darse cuatro eventos posibles de transmisión. Ahora bien, contar con información de marcadores moleculares, como así también con la fase parental, permite calcular las probabilidades de ocurrencia de cada uno de estos cuatro eventos. Asimismo, existen muchos métodos para estimar las probabilidades de transmisión. Entre ellos se encuentra el propuesto por Lander y Botstein (1989) para individuos cruza de dos líneas consanguíneas. Este procedimiento sólo toma en cuenta aquellos marcadores ubicados en las inmediaciones de los QTLs. Para el caso de poblaciones en que los apareamientos se dan al azar el cálculo de las probabilidades de trasmisión es más complejo ya que existe mayor variabilidad en el grado de información de los marcadores moleculares entre y dentro de cada familia. Con tal fin, se propusieron métodos tales como los de Liu *et al.* (2002) y Pong-Wong *et al.* (2002). Ambas metodologías estiman las probabilidades de transmisión condicionales a los marcadores cercanos a los QTLs. Otros algoritmos, tales como los de Haley *et al.* (1994) y el de Elsen *et al.* (1999), consideran todas las posibles combinaciones de alelos para cada una de las gametas. Es decir que si un padre posee $j$ marcadores heterocigotas, entonces pueden obtenerse hasta $2^j$ gametas diferentes. Estos métodos poseen la ventaja de contemplar todas las situaciones posibles que pueden darse en el proceso de transmisión genética de una generación a la siguiente. Ahora bien, desde el punto de vista operativo y computacional, es un método simple y viable siempre y cuando se emplee un número reducido de marcadores. Ahora bien, esta metodología resulta muy ineficiente al momento de trabajar con los paneles de SNPs disponibles en la actualidad con densidades de marcadores en constante crecimiento y puede tornarse totalmente inviable. Por este motivo, surgieron más recientemente métodos alternativos que permiten calcular fácilmente las probabilidades de transmisión empleando un gran número de marcadores. Uno de ellos es el propuesto por Nettelblad *et al.* (2009) que, si bien permite trabajar con densidades

elevadas de marcadores, no es eficiente a la hora de realizar los cálculos dado que toma en cuenta toda la información presente en el largo total del bloque de ligamiento. Recordemos que aquellos marcadores que pertenecen a un mismo bloque de ligamiento suelen contener información redundante ya que son loci ligados que se heredan en conjunto. Es en este aspecto donde Elsen *et al.* (2009) contribuyeron significativamente al proponer un algoritmo que lleva cabo una exploración previa de los bloques de ligamiento con el objetivo de determinar a priori el número mínimo de marcadores útiles, sin tener en cuenta información redundante. En consecuencia, este algoritmo permite trabajar eficientemente con paneles densos de marcadores para estimar las probabilidades de transmisión con requerimientos computacionales limitados en términos de tiempo, capacidad de procesamiento y memoria.

Ahora bien, para referirse específicamente al modo en que funciona el método de Elsen *et al.* (2009) es necesario definir cierta notación. Sea *p* un individuo hijo del macho *s* y la hembra *d*, todos con información genómica para *L* loci ($M_l$, $l$ = 1, 2, …, $L$). La ubicación de cada marcador ($M_l$) en el bloque de ligamiento es $x_l$ (en centi-Morgans). Se asume ausencia de interferencia en la recombinación, motivo por el cual es posible utilizar la función de distancia propuesta por Haldane (1919). La tasa de recombinación entre los loci $l_1$ y $l_2$ es igual a $r_{l1,l2}$. Si se utiliza la distancia de Haldane, $r_{l_1,l_2} = 0.5\left(1-\exp\left\{-2\left(x_{l_2}-x_{l_1}\right)\right\}\right)$. Tomando el *l*-ésimo marcador, el genotipo del padre (*s*) puede representarse como $P_{sl} = \left(P_{sl_1}, P_{sl_2}\right)$, el de la madre (*d*), como $P_{dl} = \left(P_{dl_1}, P_{dl_2}\right)$ y el del hijo (*p*), como $P_{pl} = \left(P_{pl_1}, P_{pl_2}\right)$. Para cada una de estas expresiones $\left(P_{il_k}\right)$ el subíndice *k* refiere al *k*-ésimo alelo para el marcador *l* y puede tomar valores 1 ó 2 en individuos diploides. Las probabilidades de transmisión de segmentos cromosómicos de los padres a la progenie se estiman condicionando en la información de las fases parentales. La fase de un progenitor es $G_{ilk}$ , donde *i* toma valores *s* o *d* según sea la fase paterna o materna, respectivamente, y *k* es igual a 1 ó 2 según el alelo provenga del abuelo o de la abuela, respectivamente. La fase es caracterizada por un cierto orden de los alelos $P_i = \left\{P_{ilk}\right\}$ en los loci $l$ = 1, 2, … $L$.

Ahora bien, en lo que hace al proceso de traspaso de material genético de una generación a la siguiente, sea $T\left(M_l\right)$ un evento de transmisión para el marcador *l* y $\boldsymbol{T}\left(\boldsymbol{M}\right)$ el vector de eventos de transmisión dentro de un bloque de ligamiento tal que $\boldsymbol{T}\left(\boldsymbol{M}\right)' = \left[T\left(M_1\right) \quad T\left(M_2\right) \quad ... \quad T\left(M_L\right)\right]$. Por su parte, $T\left(M_s\right)$ y $T\left(M_d\right)$ representan los eventos de transmisión del padre y de la madre a la progenie, respectivamente. Para el caso $T\left(M_{il}\right) = k$ , *k* toma valor 1 si el individuo *i* recibió el alelo del abuelo o 2 si recibió el alelo de la abuela. En consecuencia son cuatro los orígenes posibles de los alelos que recibe la progenie *p*: i) de ambos abuelos $\left(T\left(M_{sl}\right)=1, T\left(M_{dl}\right)=1\right)$ y en consecuencia $T\left(M_l\right)=11$; ii) del abuelo paterno y de la abuela materna $T\left(M_l\right)=12$ ; iii) de la abuela paterna y del abuelo materno $T\left(M_l\right)=21$; iv) de ambas abuelas $T\left(M_l\right)=22$. En el Cuadro 2.6 se

detallan las probabilidades de estos eventos de transmisión condicionales a los alelos observados para cada uno de los marcadores y a las fases parentales.

**Cuadro 2.6.** Probabilidades de los eventos de transmisión dados los genotipos observados para cada uno de los marcadores (con alelos a y b) y las fases parentales. Tomado de Elsen *et al.* (2009).

| | | | | | | $P(T(M_l) \mid G_{sl}, G_{dl}, P_{pl})$ para $T(M_l)$ = | | | |
|---|---|---|---|---|---|---|---|---|---|
| Caso | $G_{sl_1}$ | $G_{sl_2}$ | $G_{dl_1}$ | $G_{dl_2}$ | $P_{pl}$ | 11 | 12 | 21 | 22 |
| 1 | a | b | a | b | (a, a) | 1 | | | |
| 2 | a | b | a | b | (b, b) | | | | 1 |
| 3 | a | b | b | a | (a, a) | | 1 | | |
| 4 | a | b | b | a | (b, b) | | | 1 | |
| 5 | a | b | a | a | (a, a) | 1/2 | 1/2 | | |
| 6 | a | b | a | a | (a, b) o (b, a) | | | 1/2 | 1/2 |
| 7 | b | a | a | a | (a, a) | | | 1/2 | 1/2 |
| 8 | b | a | a | a | (a, b) o (b, a) | 1/2 | 1/2 | | |
| 9 | a | a | a | b | (a, a) | 1/2 | | 1/2 | |
| 10 | a | a | a | b | (a, b) o (b, a) | | 1/2 | | 1/2 |
| 11 | a | a | b | a | (a, a) | | 1/2 | | 1/2 |
| 12 | a | a | b | a | (a, b) o (b, a) | 1/2 | | 1/2 | |
| 13 | a | a | a | a | (a, a) | 1/4 | 1/4 | 1/4 | 1/4 |
| 14 | a | a | b | b | (a, b) | 1/4 | 1/4 | 1/4 | 1/4 |
| 15 | a | b | a | b | (a, b) o (b, a) | | 1/2 | 1/2 | |
| 16 | a | b | b | a | (a, b) o (b, a) | 1/2 | | | 1/2 |

$G_{il_k}$ es el alelo del marcador *l* del progenitor *i* para el cromosoma *k* (1 ó 2 según provenga del abuelo o de la abuela, respectivamente); $P_{pl}$ es el genotipo observado en la progenie; $T(M_l)$ es el evento de transmisión ocurrido en el marcador *l*; $P\left(T\left(M_l\right) \mid G_{sl}, G_{dl}, P_{pl}\right)$ es la probabilidad de que $T(M_l)$ tome valores 11, 12, 21 o 22 según las fases parentales y el genotipo observado de la progenie.

Nótese que los 16 casos presentados en el Cuadro 2.6 pueden agruparse en 5 tipos:

1) *Eventos de transmisión completamente conocidos para ambos progenitores* (casos 1 a 4). Para todos estos casos es posible inferir con probabilidad 1 el origen de ambos alelos que recibió la progenie (individuo *p*).

2) *Eventos de transmisión conocidas sólo para el padre* (casos 5 a 8). En estos casos es posible conocer con certeza el origen del alelo que el individuo *p* recibió del lado paterno, pero la incertidumbre es total del lado materno.

3) *Eventos de transmisión conocidas sólo para la madre* (casos 9 a 12). Para estos casos es posible conocer con certeza el origen del alelo que el individuo *p* recibió del lado materno, pero la incertidumbre es total del lado paterno.

4) *Eventos de transmisión desconocidos* (casos 13 y 14). En estos dos casos la incertidumbre es total tanto para la vía paterna como para la materna, motivo por el cual la probabilidad de cada evento es el valor esperado (0,25).

5) *Eventos de transmisión ambiguos* (casos 15 y 16). Estos casos corresponden a aquellos tríos de individuos (padre, madre e hijo) heterocigotas.

En las situaciones en que falta la información de alguno de los dos padres, la probabilidad condicional $T\left(M_l\right)$ que se utiliza es la del caso número 4, es decir $\boldsymbol{T}\left(M\right)' = \left[1/4 \quad 1/4 \quad 1/4 \quad 1/4\right]$. Dada la falta de información no es posible determinar con certidumbre el origen real de los alelos observados en el hijo, motivo por el cual las probabilidades asignadas corresponden a los valores esperados (0,25). De todos modos, existen casos puntuales en los que es posible llevar a cero algunas de estas probabilidades. Uno de estos casos ocurre cuando se cuenta con información para uno de los dos progenitores del individuo *p* y la fase del padre conocido es heterocigota (a, b). En tal caso, si el genotipo observado en el individuo *p* es $P_{pl} = \left(a, a\right)$, las probabilidades son $\boldsymbol{T}\left(M\right)' = \left[1/2 \quad 0 \quad 1/2 \quad 0\right]$ y, si $P_{pl} = \left(b, b\right)$, entonces $T\left(M\right)' = \left[0 \quad 1/2 \quad 0 \quad 1/2\right]$.

En consecuencia, en cualquier posición *x* del genoma, los segmentos cromosómicos recibidos por el individuo *p* pueden tener cuatro orígenes posibles: el abuelo paterno, la abuela paterna, el abuelo materno o la abuela materna. Es decir que el genotipo de un marcador del individuo *p* puede expresarse como $q = \left(q_s, q_d\right)$ donde *q* puede tomar alguno de los siguientes valores $q = \left\{\left(11\right),\left(12\right),\left(21\right),\left(22\right)\right\}$, dependiendo del origen de los alelos que recibió. En consecuencia, el método propuesto por Elsen *et al.* (2009) permite estimar la probabilidad de *q* dada la información de los marcadores y de las fases parentales, es decir $P_x\left(q\right) = P\left[T\left(Q_x\right) = q \mid G_s, G_d, G_p\right]$. Con tal propósito, se emplean dos propiedades importantes. Por una parte, los eventos de transmisión del padre y de la madre son independientes marginalmente al genotipo de los marcadores. Es decir, $P\left[T\left(M_l\right)\right] = P\left[T\left(M_{sl}\right)\right]P\left[T\left(M_{dl}\right)\right]$. Además, dado el supuesto de ausencia de interferencia, los eventos de transmisión siguen un proceso markoviano, descripto por la siguiente expresión:

$$P\left[T\left(M\right)\right] = P\left[T\left(M_1\right)\right]P\left[T\left(M_2\right) \mid T\left(M_1\right)\right]P\left[T\left(M_3\right) \mid T\left(M_2\right)\right]...P\left[T\left(M_L\right) \mid T\left(M_{L-1}\right)\right] \quad [24]$$

Valiéndose de estas dos propiedades, el método propuesto por Elsen *et al.* (2009) permite computar las probabilidades de transmisión descriptas de modo eficiente en términos computacionales. A tal fin involucra dos aspectos centrales: una exploración iterativa de los bloques de ligamiento y una reducción de cada grupo de ligamiento que permite tomar en cuenta aquella información relevante a los efectos de los cálculos de las probabilidades de interés. Ambos aspectos permiten tomar los casos realmente informativos dentro de cada bloque de ligamiento. Esto permite trabajar con un subgrupo de marcadores verdaderamente informativos para calcular las probabilidades de transmisión de todo el

bloque, dejando de lado casos tales como los de transmisión desconocida o conocida sólo para alguno de ambos padres. Así es posible estimar las probabilidades de transmisión de padres a hijos utilizando información mínima, y con requerimientos computacionales muy bajos tanto en tiempo como en memoria y capacidad de procesamiento. Un aspecto a destacar de este método es que no requiere frecuencias alélicas ya que en todo momento los cálculos se efectúan condicionalmente a los alelos observados en los individuos y a las fases parentales.

## 2.5. El método de VanRaden (2008)

VanRaden (2007, 2008) propuso calcular las relaciones genómicas realizadas considerando solamente la información de los marcadores moleculares sin tener en cuenta la información genealógica. Para ello sugirió calcular la matriz $\boldsymbol{G}$ de relaciones genómicas sobre la base del producto cruzado de los genotipos en cada locus, divididos por la heterocigosidad total de cada marcador. Las relaciones genómicas resultantes reflejan la proporción de genoma IBS que comparten dos individuos, expresada como una desviación de la proporción esperada de alelos compartidos en la población (Vela-Avitua *et al.*, 2015). Operacionalmente, se define una matriz $\boldsymbol{M}$ de orden $n$ (número de individuos) por $m$ (número de marcadores por individuo). Cada uno de sus elementos indica el número de copias del alelo de referencia qué posee cada individuo en cada marcador y puede tomar uno de tres posibles valores: (i) 0 cuando el genotipo del individuo $i$ para el marcador $j$ es homocigota 11; (ii) 1 cuando el individuo $i$ es heterocigota 12 o 21 para el SNP $j$; (iii) 2 cuando el individuo es homocigota 22 para el marcador $j$. Cabe destacar que el método trabaja bajo los supuestos de equilibrio Hardy-Weinberg (EHW) y equilibrio de ligamiento (LE) entre marcadores. Por otro lado, se construye una matriz $\boldsymbol{P}$ ($n$ x $m$) que contiene las frecuencias alélicas expresadas como la diferencia de 0,5 multiplicada por dos, de manera tal que la columna $i$ de $\boldsymbol{P}$ es 2 ($p_i - 0{,}5$) donde $p_i$ es la frecuencia del alelo de referencia en el locus $i$. Una vez calculadas ambas matrices se obtiene la matriz $\boldsymbol{Z}$ del siguiente modo:

$$\boldsymbol{Z} = \boldsymbol{M} - \boldsymbol{P} \qquad [25]$$

Al momento de calcular las relaciones genómicas, el hecho de restar las frecuencias alélicas presentes en $\boldsymbol{P}$ a la información de los genotipos (en $\boldsymbol{M}$) permite dar mayor importancia a aquellos alelos con baja frecuencia (en general altamente informativos del parentesco entre individuos) en la población. Las frecuencias alélicas en $\boldsymbol{P}$ deben ser de la población base no selecta. Utilizar aquellas frecuencias alélicas calculadas utilizando la información molecular de individuos de generaciones recientes puede dar lugar a sesgos (VanRaden, 2008).

El primer método descripto por VanRaden (2008) se basa en la siguiente expresión para computar ***G***

$$\boldsymbol{G} = \frac{\boldsymbol{Z}\,\boldsymbol{Z}'}{2\sum p_j\left(1-p_j\right)} \qquad [26]$$

El hecho de dividir por $2\sum p_j\left(1-p_j\right)$ permite escalar ***G*** para que sea análoga a la matriz ***A*** de relaciones aditivas. La matriz ***G*** del primer método propuesto por VanRaden es positiva semidefinida, pero puede ser singular en casos en que dos o más individuos posean genotipos idénticos como es el caso de los clones. ***G*** será singular cada vez que el número de marcadores sea inferior al de animales ($m < n$). Estas propiedades de ***G*** tienen gran impacto al momento de utilizar dicha matriz para obtener predicciones del mérito genético de los individuos.

## 2.6. Comparación de metodologías para la estimación de las relaciones de parentesco realizadas.

En el Cuadro 2.7 se resumen los principales aspectos de cada uno de los métodos descriptos en las secciones anteriores. El objetivo de la misma es comparar muy brevemente los métodos entre sí con el objetivo de notar sus similitudes y diferencias en ciertos aspectos claves. De dicho cuadro se desprende que este subconjunto de cuatro métodos pretende representar la vasta variedad de algoritmos disponibles para la estimación de las GWR.

**Cuadro 2.7.** Comparación de las cuatro metodologías descriptas para la estimación de relaciones de parentesco realizadas.

| | Método | | | |
|---|---|---|---|---|
| Aspecto | **Guo** | **Han y Abney** | **Elsen *et al.*** | **VanRaden** |
| Necesita contar con la fase | SI | NO | SI | NO |
| Utiliza información de pedigrí | SI | SI | SI | NO |
| Utiliza información de marcadores | SI | SI | SI | SI |
| Tiene en cuenta la consanguinidad | NO | SI | NO | SI |
| Tiene en cuenta el LD | NO | SI | Considera el bloque de ligamiento | NO |
| Metodología subyacente | MM | HMM | MM | Producto cruzado de genotipos |
| Aplicabilidad | Sólo MH | Pedigríes completos | Pedigríes muy sencillos | Pedigríes completos |
| Año en que fue presentado | 1994 | 2011 | 2009 | 2007-2008 |

Nótese que en todos los casos se utiliza la información de los marcadores moleculares pero ocurre lo mismo con la información genealógica. La misma es considerada por los métodos de Guo, Han y Abney y Elsen *et al.*, pero no en el método de VanRaden que emplea estadísticos poblacionales como la información de las frecuencias alélicas.

## 2.7. Varianza de las relaciones de parentesco realizadas

Existe cierta variabilidad en el valor que puede tomar una GWR entre dos individuos. Esto se debe a la estocasticidad de procesos como la segregación mendeliana y la recombinación al momento de la formación de las gametas. Tómese por ejemplo el caso de los hermanos enteros. Si sólo se utiliza información genealógica, no se dispone de información sobre lo sucedido durante la segregación, motivo por el cual el cálculo de la relación entre hermanos enteros en esta situación (y ausencia de consanguinidad) siempre produce 0,5 (valor esperado). Ahora bien, el resultado cambia al considerar además información genómica en los cálculos. Esto se debe a que los marcadores moleculares

permiten recuperar información respecto a lo ocurrido durante la segregación. Por ejemplo, Gagnon *et al.* (2005) reportaron que el promedio de la GWR entre hermanos enteros fue 0,4994 con un desvío estándar de 0,0395. Nótese que si se asume una distribución normal, el resultado implica que el 99% de los pares de individuos poseen una relación de parentesco que varía entre 0,3809 y 0,6179. En consecuencia, es notable la importancia de conocer la varianza asociada a las GWR. La selección genómica depende de la existencia de variabilidad en la proporción de genoma compartido entre individuos emparentados cuya relación de parentesco por pedigrí es la misma (Meuwissen *et al.*, 2001). Durante varios años se llevaron a cabo grandes esfuerzos para cuantificar la varianza de diferentes relaciones de parentesco y para generar algoritmos de cálculo. Risch y Lange (1979) fueron los primeros en presentar una expresión teórica para calcula la varianza de la GWR entre hermanos enteros, tomando en cuenta la información de todo el genoma. Por su parte Suarez *et al.* (1979) utilizaron simulaciones para estimar la variabilidad de PIBD entre pares de hermanos. Posteriormente, Rasmuson (1993) obtuvo una fórmula para calcular la variabilidad de PIBD, utilizando el *índice de recombinación*, útil para diferentes relaciones de parentesco. El *índice de recombinación* fue definido por Darlington (1939) como el número haploide de cromosomas más el total de quiasmas (puntos de inserción o de unión entre las cromátidas hermanas que forman el par cromosomas homólogos). Si bien este resultó ser un enfoque pragmático para obtener estimaciones de la variabilidad, supone al genoma como un conjunto de unidades discretas independientes. Tal como demostraron Donnelly (1983) y Thompson (1993), dicho supuesto no es realista dado que no es posible convertir el genoma continuo a un número equivalente de loci que segregan de modo independiente. Hill (1993 a, b) y Guo (1995, 1996) presentaron fórmulas para calcular la varianza teórica esperada de la PIBD para ciertas relaciones de parentesco, tales como ancestro-descendiente, medio-hermanos y primos. En trabajos posteriores, Hill y Weir (2011) presentaron un marco teórico que permitió unificar diversas aproximaciones presentadas con anterioridad para el cálculo de las varianzas de ciertas relaciones de parentesco. En dicho trabajo, se presentaron fórmulas para diferentes tipos de relaciones de parentesco, sobre el supuesto de la inexistencia de consanguinidad. Esta propuesta permite el cálculo de la varianza en relaciones de parentesco ya sean cercanas o lejanas, al emplear las probabilidades de compartir 0, 1 o 2 pares de alelos IBD representadas por $k_0$, $k_1$ y $k_2$, respectivamente. Hill y Weir (2011) cuantificaron también en qué magnitud difiere la varianza de relaciones de parentesco cuyo valor esperado es el mismo. A modo de ejemplo se puede tomar el caso de bisabuelo - bisnieto, tío abuelo - sobrino nieto, medio tío – sobrino y primos hermanos. Nótese que en todos estos casos la relación de parentesco esperada es de 0,125, mientras que sus desvíos estándar varían, aspecto fundamental a la hora de distinguir entre las diferentes relaciones. En un trabajo posterior Hill y Weir (2012) extendieron dichas fórmulas para incluir los casos con consanguinidad. Demostraron que la consanguinidad en uno de los padres impacta a nivel de la varianza de la relación de parentesco observada y tanto su magnitud como efecto dependen no sólo del nivel de consanguinidad sino también del tipo de relación de parentesco evaluada (Hill y Weir, 2012). García-Cortés *et al.* (2013) presentaron una fórmula general aplicable a todas las relaciones de parentesco que permite calcular las varianzas en términos de los nueve estados condensados de identidad desarrollados por Harris (1964) y Gillois (1964) (posteriormente empleados por Jacquard en 1974) y los coeficientes de coancestría presentados por Karigl (1981).

Se presenta a continuación la expresión que permite calcular la varianza teórica esperada para medio-hermanos siguiendo la derivación presentada por Hill (1993 a), Visscher *et al.* (2006) y Visscher (2009). La relación de parentesco esperada para medio-hermanos es de 0,25. Al considerar la información molecular, es posible calcular la proporción de genoma compartido IBD real u observado, o lo que es equivalente, calcular la relación de parentesco genómica que denotaremos como $\pi$. Esta es una variable aleatoria cuya esperanza coincide con la relación de parentesco esperada ($E(\pi) = A_{ij}$). Ahora bien, a continuación se detalla cómo calcular $Var(\pi)$ según se consideren uno, dos o varios loci en el cálculo de $\pi$. Para un solo locus, sea $\delta_i$ una variable indicadora para el locus *i*. La misma es igual a 1 cuando ambos individuos de la pareja de medio-hermanos portan el mismo alelo proveniente del progenitor en común para el locus *i*, o es igual a 0 en todos los otros casos. Para medio-hermanos $P(\delta_i = 1) = P(\delta_i = 0) = 0,5$ y, en consecuencia, $\pi_i = 0,5\delta_i$. Por lo tanto, $E(\pi_i) = 0,25$ y $Var(\pi_i) = 0,25$. La situación difiere si se considerasen dos loci *i* y *j* con una fracción de recombinación *c*, donde

$$E\left(\pi_i, \pi_j\right) = 0,0625\left[2\left(1-c\right)^2 + 2c^2\right] \qquad [27]$$

$$Cov\left(\pi_i, \pi_j\right) = E\left(\pi_i, \pi_j\right) - E\left(\pi_i\right)E\left(\pi_j\right) = 0.0625\left(1-2c\right)^2 \qquad [28]$$

Si los eventos de recombinación se asumen distribuidos uniformemente y sin interferencia tal como lo propuso Haldane (1919), la covarianza puede calcularse del siguiente modo

$$Cov\left(\pi_i, \pi_j\right) = 0,0625\exp\left\{-4\,d_{ij}\right\} \qquad [29]$$

donde $d_{ij}$ es la distancia entre los loci *i* y *j*, medida en Morgans.

Si se generaliza a *n* loci, la variancia para la relación de medio-hermanos es igual a

$$\mathrm{Var}\left(\pi\right) = \left(\tfrac{1}{n^2}\right)0,0625\sum\sum\exp\left\{-4\,d_{ij}\right\} \qquad [30]$$

Nótese que si *n* aumenta considerablemente, la ecuación [30] puede expresarse como una integral (Hill, 1993 a; Stam y Zeven, 1981):

$$\mathrm{Var}\left(\pi\right) = \frac{1}{16}\int_0^l\int_0^l e^{-4|x_1-x_2|}\,dx_1\,dx_2 = \frac{1}{32l^2}\left(l - \frac{r_{2l}}{2}\right) \qquad [31]$$

El valor *l* es el largo del cromosoma medido en Morgans y $r_{2l}$ es la fracción de recombinación para un segmento de largo *2l* ($0,5\,(1 - \exp\{-4l\})$. En consecuencia, la varianza de la relación observada entre dos medio-hermanos para un cromosoma de largo *l* es (Guo 1996, Hill 1993 a):

$$\mathrm{Var}(\pi) = \frac{1}{128 l^2}\left[4l - 1 + \exp\{-4l\}\right] \qquad [32]$$

Ahora bien, es posible calcular una $\pi$ global para todo el genoma compuesto de $k$ cromosomas como un promedio ponderado de las $\pi_i$ de cada cromosoma ($i = 1, \ldots, k$). A tal fin se emplea la siguiente expresión:

$$\pi = \frac{1}{L}\sum l_i \, \pi_i \qquad [33]$$

El valor $L$ representa el largo total del genoma completo y es calculado como $L = \sum l_i$ . En consecuencia, se puede generalizar la expresión [32] para un genoma compuesto por $k$ cromosomas que segregan independientemente del siguiente modo:

$$\mathrm{Var}(\pi) = \frac{1}{128 L^2}\left[4L - k + \sum_i \exp\{-4 l_i\}\right] \qquad [34]$$

# Capítulo 3. *Materiales y Métodos*

# Capítulo 3

# Materiales y Métodos

## 3.1. Animales

La base de datos utilizada para evaluar y comparar el desempeño de los cuatro métodos de estimación proviene de una población experimental de 411 cerdos distribuidos en tres generaciones ($F_0$, $F_1$ y $F_2$) y criados en Michigan State University Swine Teaching and Research Farm, East Lansing, Michigan (Edwards *et al.*, 2008). En la generación inicial ($F_0$) se cruzaron por inseminación artificial 15 hembras de la raza Pietrain con cuatro machos de la raza Duroc. Estos servicios produjeron los individuos $F_1$ de los cuales se seleccionaron 50 hembras y seis machos como padres de la próxima generación. Tanto en la $F_0$ como en la $F_1$ se seleccionaron animales no relacionados entre sí con el objetivo de evitar la consanguinidad en las siguientes generaciones. Finalmente, en la tercera generación ($F_2$) se seleccionaron 336 individuos de un total de 1259 lechones nacidos en 141 camadas. Se registraron en total, 6704 pares de medio-hermanos $F_2$ disponibles para el estudio.

## 3.2. Genotipado y Edición de Datos Genómicos

Los 411 animales de la población experimental fueron genotipados utilizando el chip Illumina PorcineSNP60 beadchip (Ramos *et al.*, 2009) en un laboratorio comercial (GeneSeek, Neogen Company, Lincoln, NE, USA). Este panel cuenta con un total de 62163 SNPs, de los cuales 13970 fueron eliminados en una primera instancia debido a que su ubicación en el genoma era desconocida. Además, se consideraron solamente aquellos marcadores autosómicos, motivo por el cual 1328 marcadores ubicados en los cromosomas X e Y no fueron incluidos en el análisis. Se buscaron posibles inconsistencias mendelianas, tanto nivel de SNP como de individuo, utilizando el software PLINK v1.07 (Purcell *et al.*, 2007). Una vez detectadas dichas inconsistencias, se les asignó el código de genotipo faltante a los marcadores involucrados. Finalmente, se fijaron dos valores mínimos del alelo en menor frecuencia (MAF) para generar dos conjuntos de datos genómicos distintos en términos de la densidad de marcadores y de la informatividad de los mismos. Por un lado, se fijó un valor de MAF de 0,01, eliminando aquellos marcadores que presentaron valores inferiores. Un valor de MAF de 0,01 es comúnmente utilizado a la hora de editar bases de datos genómicos, tal como se reporta frecuentemente en la literatura. Con este valor umbral se busca eliminar aquellos SNPs en los que se observa una variante alélica distinta en el 1% de los casos, normalmente asociado a errores de genotipado. Por otra parte, en lo que puede considerarse una situación “extrema”, el umbral de MAF se fijó en 0,20. Se califica a este escenario como extremo dado que con valores tan elevados de MAF se pierden muchos SNPs durante el proceso de edición de la base de datos genómicos. El objetivo de fijar dicho valor umbral elevado fue generar cambios importantes en términos de cantidad y de

información de los marcadores que permitiesen evaluar la sensibilidad de los métodos descriptos. Como consecuencia, aplicando el primer umbral de MAF se eliminaron 7743 marcadores cuya frecuencia del alelo menos frecuente se encontraba por debajo de 0,01. De este modo se obtuvo el primer conjunto de datos genómicos denominado DMAF_0,01, con los genotipos de 411 animales cada uno con 39122 marcadores. Por otro lado, al emplear el segundo umbral de MAF, se generó un segundo conjunto de datos denominado DMAF_0,20 con el mismo total de animales, pero con 25957 SNPs por individuo. Previo a utilizar ambas bases para estimar las GWR, se verificó que los marcadores remanentes post-edición mantuvieran el patrón de distribución original (en términos de proporción del total de marcadores por cromosoma) a lo largo del genoma tal como se observa en la Figura 3.1. Nótese que el hecho de disminuir la densidad de marcadores como consecuencia de aumentar el valor umbral de MAF en la edición, no cambió significativamente el patrón de cobertura original del genoma.

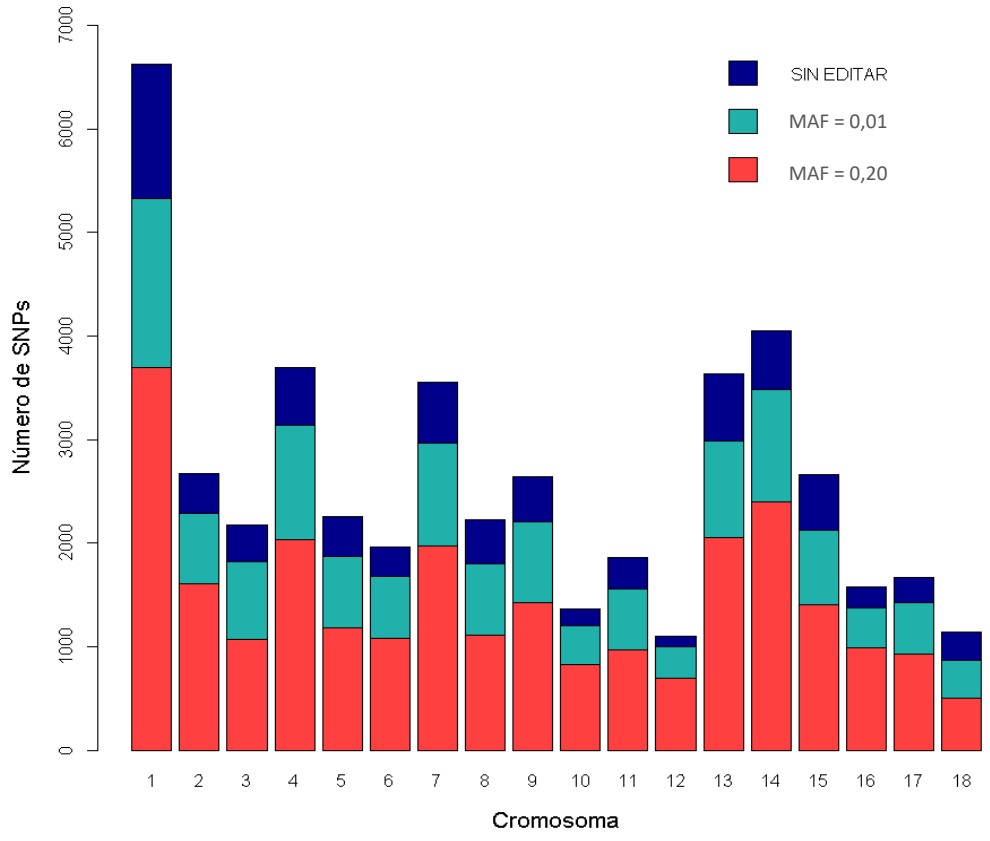



**Figura 3.1.** Distribución de los marcadores SNPs a lo largo del genoma autosómico compuesto por 18 cromosomas. Se contemplan tres conjuntos de marcadores: i) *SIN EDITAR* corresponde al conjunto inicial, previo a la edición por MAF; ii) *MAF = 0,01* corresponde al conjunto de datos DMAF_0,01 (filtrados empleando un MAF umbral de 0,01); iii) *MAF = 0,20* corresponde al conjunto de datos DMAF_0,20 (filtrados empleando un MAF umbral de 0,20).

### 3.3. Obtención de las fases parentales

Tanto el método de Guo (1994) como el de Elsen *et al.* (2009) necesitan contar con las fases parentales conocidas. Estas fueron estimadas según el método de Browning y Browning (2009) (implementado en el software Beagle 3.3.2) y aquel propuesto por Favier *et al.* (2010) (implementado dentro del software QTLMap). El último permite estimar también las probabilidades de transmisión de segmentos cromosómicos de padres a hijos siguiendo el método de Elsen *et al.* (2009) descripto anteriormente.

El programa Beagle 3.3.2 (Browning y Browning, 2009) permite inferir las fases más probables mediante un algoritmo HMM, condicionando en los genotipos observados de los padres y de la progenie (Browning y Browning, 2007). Se utilizó la opción *trio* que permite ingresar la información genómica por tripletes de individuos. De este modo, para los medio-hermanos de la población experimental de cerdos primero se ingresaron los datos genómicos de ambos padres y luego los del hijo. Por otra parte, para estimar las GWR entre medio-hermanos empleando el método de Elsen *et al.* (2009) fue necesario obtener las fases parentales con el algoritmo de Favier *et al.* (2010). Este método permite reconstruir las fases por máxima verosimilitud, empleando un algoritmo de búsqueda eficiente. Fue desarrollado para obtener las fases parentales de familias de medio-hermanos tomando la información genómica parental y de la progenie.

### 3.4. Estimación de las relaciones de parentesco observadas

Las GWR de los medio-hermanos en la población experimental de cerdos fueron calculadas mediante cuatro métodos: (i) el de Guo (1994), (ii) el de Han y Abney (2011), (iii) el de Elsen *et al.* (2009) y (iv) el de VanRaden (2008). Las principales diferencias entre ellos se resumen en el Cuadro 2.7. Además, para cada metodología se obtuvieron estimaciones con ambos conjuntos de datos genómicos filtrados por MAF: DMAF_0,01 y DMAF_0,20. En consecuencia las GWR de las 6704 parejas de medio-hermanos en la generación $F_2$ se estimaron ocho veces en un diseño factorial $2 \times 4$: dos valores umbrales de MAF empleados para editar los datos genómicos por cuatro métodos de estimación.

#### 3.4.1 El método de Guo (1994)

Para estimar la PIBD entre pares de medio-hermanos en la población experimental de cerdos se desarrolló un programa Fortran con el objetivo de implementar la metodología propuesta por Guo (1994). El programa consta de tres secciones principales:

I. Exploración de cada segmento cromosómico que permite determinar qué caso de los presentados en el Cuadro 2.1 se observa en cada segmento. A tal efecto, se utiliza la información de las fases parentales y de los genotipos observados para cada individuo. El programa permite comparar los alelos del progenitor con los de ambos medio-hermanos para el par de marcadores que delimitan un segmento.

II. Cálculo de la PIBD en cada segmento empleando la expresión apropiada del Cuadro 2.1 para cada caso particular a lo largo del genoma.

III. Calculo de la PIBD global mediante la fórmula [1] para cada par de medio-hermanos. Posteriormente, la GWR para cada par de medio-hermanos se calcula como dos veces el valor de PIBD global.

### 3.4.2 El método de Han y Abney (2011)

Se utilizó el programa IBDLD v3.13 (Han y Abney, 2013) para estimar las GWR siguiendo el método de Han y Abney (2011). El programa emplea las probabilidades de los nueve modos de identidad condensados presentados por Jacquard (1974) y descriptos en el Cuadro 2.3. Para calcularlas se empleó el programa "IdCoefs 2.1" (Abney, 2009). En esta etapa inicial sólo se utilizó la información de la genealogía para el cálculo. Posteriormente, las probabilidades fueron actualizadas condicionando en la información de los marcadores moleculares, tal como lo proponen Han y Abney (2011) y cómo fue descripto en la sección 2.3.2.2. Como resultado se estimaron los valores de la PIBD compartido entre cada par de medio-hermanos (*x* e *y*), para cada uno de los 18 cromosomas autosómicos. A los efectos de obtener una estimación global para todo el genoma ($PIBD_{xy(global)}$), se calculó un promedio ponderado de las PIBD cromosómicas ($PIBD_i$), considerando el largo ($c_i$) de cada uno de los *N* cromosomas según la siguiente expresión

$$PIBD_{xy(global)} = \frac{\sum_{i=1}^{N} c_i \, PIBD_i}{L} \qquad [35]$$

Donde *L* es la longitud de todo el genoma autosómico que, para el caso del cerdo, corresponde a la suma de las longitudes de los 18 cromosomas autosómicos (*N* = 18), es decir $L = \sum_{i=1}^{18} (c_i)$.

### 3.4.3 El Método de Elsen *et al.* (2009)

La estimación de las GWR entre medio-hermanos de la población experimental requirió de dos etapas de cálculo. En la primera se procedió a estimar las probabilidades de transmisión de segmentos cromosómicos de padres a hijos para cada posición del genoma según Elsen *et al.* (2009). A tal fin se utilizó el programa QTLMap desarrollado por Filangi *et al.* (2010). En un paso inicial, el programa permite obtener las fases parentales empleando el algoritmo de Favier *et al.* (2010). Para estimar las probabilidades de transmisión, el programa de Filangi *et al.* (2010) admite una amplia gama de opciones para realizar diferentes ajustes, tal como se detalla en el manual. Para los medio-hermanos de la población experimental se utilizó la opción *snp* para aumentar la velocidad de cálculo. Dicha opción permite lograr velocidades de cómputo significativamente más altas al

trabajar con paneles densos de marcadores, como es el caso de DMAF_0,01 y DMAF_0,20. También se empleó la opción *opt_step 0* para realizar el análisis en la posición de cada marcador. Una vez calculadas las probabilidades de transmisión, en una segunda etapa, se estiman las PIBD entre medio-hermanos. Se presenta a continuación los pasos para calcular las PIBD. Los mismos fueron implementados en un programa escrito en lenguaje *R*.

Las relaciones de parentesco realizadas empleando probabilidades de transmisión se calculan bajo el supuesto que los individuos analizados son medio-hermanos. Dada esta condición, cada pareja de individuos puede compartir sólo un cromosoma parental, sea el materno o el paterno, dependiendo si están relacionados por vía materna (medio-hermanos maternos) o por vía paterna (medio-hermanos paternos), respectivamente. Al evaluar un segmento cromosómico de un par de medio-hermanos paternos (individuo *i* y *j*) pueden observarse cuatro situaciones posibles, tal como se esquematiza en la Figura 3.2: (i) tanto *i* como *j* recibieron el segmento cromosómico proveniente del abuelo paterno de ambos; (ii) *i* recibió el segmento cromosómico de su abuelo paterno mientras que *j* recibió el de su abuela paterna; (iii) *i* recibió el segmento cromosómico de su abuela paterna mientras que *j* recibió el de su abuelo paterno; (iv) ambos *i* y *j* recibieron el segmento cromosómico por parte de la abuela materna. Ahora bien, dado que para ser IBD ambos segmentos cromosómicos deben ser copia de uno ancestral, sólo las situaciones (i) y (iv) cumplen con esta condición. En consecuencia, la probabilidad de IBD en una posición dada puede expresarse como:

$$P\left(IBD_{ij}\right) = P\left(i \leftarrow s\right)P\left(j \leftarrow s\right) + P\left(i \leftarrow d\right)P\left(j \leftarrow d\right) \qquad [36]$$

donde $P\left(i \leftarrow s\right)$ representa la probabilidad de que el individuo *i* haya recibido el segmento cromosómico paterno (del abuelo paterno), $P\left(j \leftarrow s\right)$ es la probabilidad de que el individuo *j* haya recibido el segmento cromosómico paterno (del abuelo paterno), $P\left(i \leftarrow d\right)$ representa la probabilidad de que el individuo *i* haya recibido el segmento cromosómico materno (de la abuela paterna) y $P\left(j \leftarrow d\right)$, la probabilidad de que el individuo *j* haya recibido el cromosoma materno (de la abuela paterna). Dado que los individuos bajo estudio son diploides (cuentan con dos juegos de cromosomas) y que la probabilidad de que el animal *i* haya recibido el segmento cromosómico del abuelo paterno es $p_i$ ($P\left(i \leftarrow s\right) = p_i$), entonces la probabilidad de haber recibido el segmento cromosómico de la abuela paterna debe ser $1 - p_i$ ($P\left(i \leftarrow d\right) = 1 - p_i$). Consecuentemente, para un segmento específico, la probabilidad de IBD entre los individuos *i* y *j* es:

$$P\left(IBD_{ij}\right) = p_i p_j + \left(1 - p_i\right)\left(1 - p_j\right) \qquad [37]$$

Para obtener una estimación global (*genome-wide*) de la proporción del genoma IBD, basado en las estimaciones para cada segmento, se utiliza la siguiente fórmula:

$$PIBD_{ij} = \frac{1}{n}\left[\boldsymbol{p}_i' \boldsymbol{p}_j + \left(\mathbf{1} - \boldsymbol{p}\right)_i' \left(\mathbf{1} - \boldsymbol{p}\right)_j\right] \qquad [38]$$

donde $\boldsymbol{p}_i$ representa un vector cuyos elementos son las probabilidades de que el individuo *i* haya recibido el segmento cromosómico del abuelo paterno, para cada ubicación del cromosoma. Análogamente, $(\mathbf{1}-\boldsymbol{p})_i$ es un vector con las probabilidades de que el individuo *i* haya recibido el segmento cromosómico de la abuela paterna para cada posición del genoma; *n* es el número total de segmentos cromosómicos considerados. Nuevamente, para calcular la GWR para cada par de medio-hermanos se procede a multiplicar por dos la PIBD global.

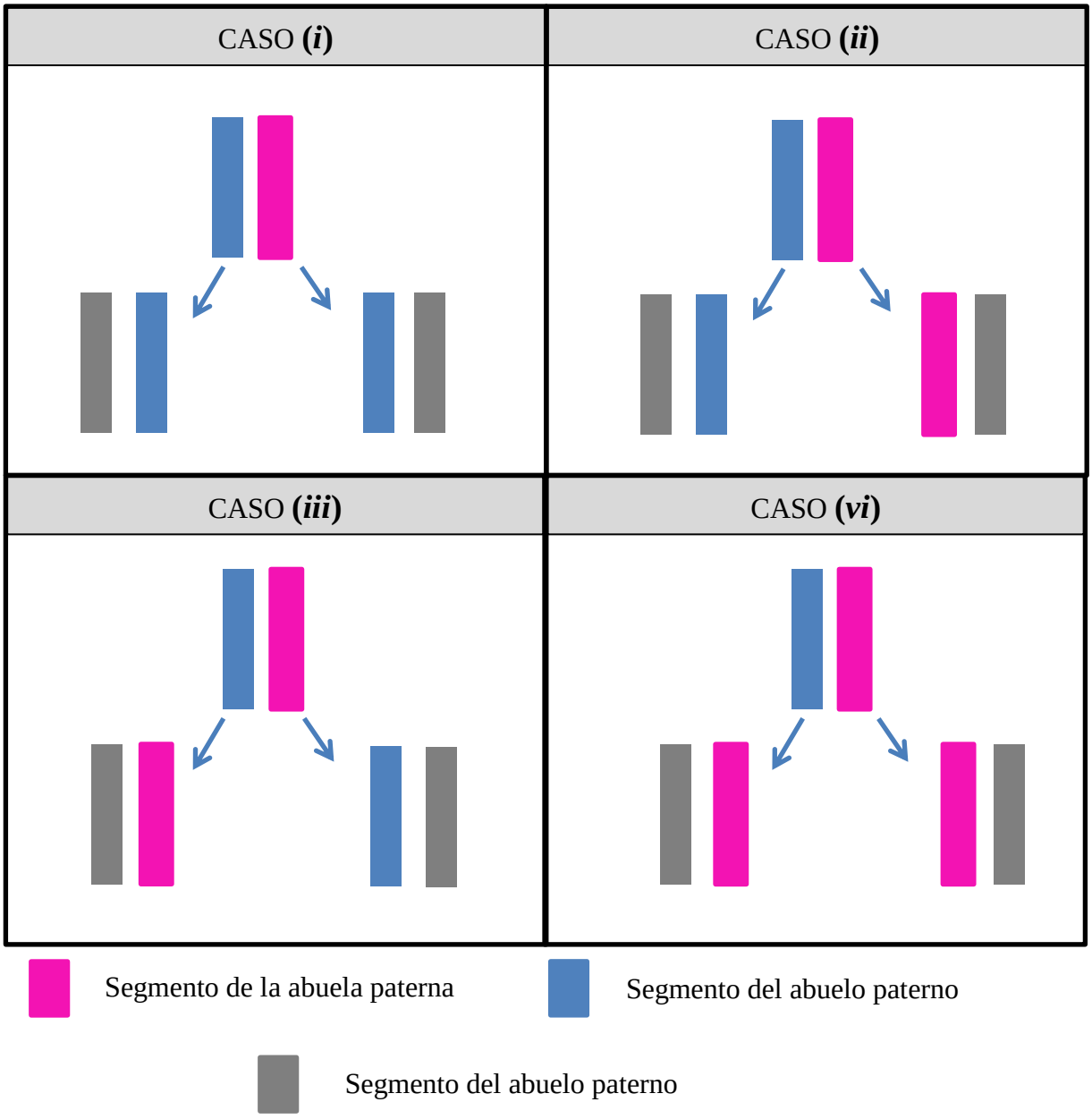



**Figura 3.2.** Ejemplo para un segmento cromosómico en el que se detallan los cuatro eventos posibles de transmisión que pueden darse para el caso de los medio-hermanos paternos.

Para ilustrar cómo se calcula la PIBD compartida entre dos individuos medio-hermanos a partir de las probabilidades de transmisión según Elsen *et al.* (2009), considerese el ejemplo en la Figura 2.2 (sección 2.2). El vector $\boldsymbol{p}_1$ contiene las probabilidades de que el individuo 1 haya recibido el segmento cromosómico paterno en cada ubicación cromosómica. Dado que el individuo 1 recibió el segmento completo del padre, $\boldsymbol{p}_1$ es igual a $\boldsymbol{p}_1' = \begin{bmatrix} 1 & 1 & 1 & 1 \end{bmatrix}$. Análogamente, $\boldsymbol{p}_2$ contiene las probabilidades de que el individuo 2 haya recibido el segmento cromosómico paterno en cada ubicación cromosómica, y es igual a $\boldsymbol{p}_2' = \begin{bmatrix} 1 & 1 & 0{,}5 & 0 \end{bmatrix}$. Consecuentemente, $\left(\mathbf{1} - \boldsymbol{p}_1\right)' = \begin{bmatrix} 0 & 0 & 0 & 0 \end{bmatrix}$ y $\left(\mathbf{1} - \boldsymbol{p}_2\right)' = \begin{bmatrix} 0 & 0 & 0{,}5 & 1 \end{bmatrix}$. Para estimar la proporción del genoma IBD en el segmento entero se emplea [38] de modo tal que

$$\text{PIBD}_{12} = 0{,}125\left[\boldsymbol{p}_i'\boldsymbol{p}_j + \left(\mathbf{1} - \boldsymbol{p}\right)_i' \left(\mathbf{1} - \boldsymbol{p}\right)_j\right] = 0{,}125\left[2{,}5 + 0\right] = 0{,}31$$

Nuevamente es necesario dividir por ocho ($n$ = 8) dado que para determinar qué proporción del segmento cromosómico comparten los individuos, no sólo es necesario considerar los cuatro segmentos de origen paterno sino también los cuatro provenientes del lado materno para la misma región cromosómica donde, en todos los casos, la probabilidad que sean IBD es 0. De la estimación de PIBD se desprende entonces que los individuos 1 y 2 comparten aproximadamente un 31% de la región cromosómica analizada. Nótese que este valor coincide con aquel obtenido según el método de Guo (1994) con los mismos datos.

### 3.4.4 El Método de VanRaden (2008)

Las relaciones de parentesco observadas también fueron estimadas con el primer método propuesto por VanRaden (2008) descripto en la sección 2.5. Recordemos que esta metodología solo utiliza la información de los marcadores moleculares, sin considerar la genealogía. Para estimar dichas relaciones se utilizó el programa preGSf90 (Aguilar *et al.*, 2011), empleando la opción *tunedG 0* para calcular la matriz $\boldsymbol{G}$ original de VanRaden (2008) sin escalarla o combinarla con la matriz $\boldsymbol{A}$. El programa suele realizar este último paso por defecto para asegurar que la matriz de relaciones genómicas resultante sea positiva definida. Para el cálculo de las GWR se utilizaron las frecuencias alélicas de la generación $F_0$. Como se mencionó anteriormente, los genotipos provinieron de una población de cerdos cuya generación inicial tenía 19 animales, cuatros padres Duroc y 15 hembras Pietrain. Para tomar en cuenta las diferencias existentes a nivel de frecuencias alélicas entre ambas razas y su consecuente contribución diferencial al *pool* de gametas fundadoras, se empleó un promedio ponderado de los valores de ambas razas. Como resultado se obtuvo la matriz $\boldsymbol{G}$ de relaciones genómicas para los individuos de la población experimental. En consecuencia, fue necesario extraer aquellos elementos de $\boldsymbol{G}$ que correspondían a las GWR entre medio-hermanos de la generación $F_2$.

### 3.5. Cálculo de la varianza teórica esperada según Hill (1993)

Una vez estimadas las GWR con cada metodología para cada una de las 6704 parejas de medio-hermanos, se procedió a evaluar el desempeño de cada uno de los métodos y a comparar los resultados entre sí. Se evaluó la varianza empírica de las relaciones estimadas captada por cada método en relación a la esperada según la teoría. Esta última se calculó con la expresión propuesta por Hill (1993 a) y presentada en [34]. La misma depende del número de cromosomas ($k$), de la longitud de cada uno ($l_i$) y de la longitud del genoma completo ($L$). Para los cálculos se emplearon los valores del mapa genético porcino presentado por Tortereau *et al.* (2012): $k$ = 18, una longitud cromosómica promedio de 0,92 Morgans y el largo total del genoma de 16,56 Morgans. Cabe aclarar que si bien se reporta el valor promedio para la longitud de los cromosomas, al momento de llevar a cabo los cálculos se tomó la longitud de cada cromosoma según lo descripto por Tortereau *et al.* (2012).

### 3.6. Análisis estadístico

Al momento de analizar y comparar las estimaciones de las GWR obtenidas por las cuatro metodologías para ambos conjuntos de datos, se efectuó la prueba de Levene que mostró evidencias de heterogeneidad de varianzas entre métodos. Como consecuencia se ajustó un modelo de una vía (tomando como tratamiento cada combinación de método de estimación y de conjunto de datos empleado) teniendo en cuenta la heterocedasticidad con una estructura diagonal en bandas para la matriz de covarianzas asociadas al tratamiento. Para llevar a cabo estos análisis estadísticos se utilizó PROC MIXED de SAS (SAS v.9.3.1; SAS Institute Inc., Cary NC, USA), empleando la corrección de los grados de libertad por el procedimiento propuesto por Kenward y Roger (1997). Finalmente, se llevaron a cabo pruebas de comparaciones múltiples entre métodos empleando el procedimiento de Tukey. En todos los casos se consideraron significativos los valores del estadístico asociados con probabilidades menores o iguales a 0,05.

# Capítulo 4. *Resultados*

# Capítulo 4

# Resultados

## 4.1. Desempeño de los métodos de estimación de las relaciones de parentesco realizadas

En el Cuadro 4.1 se presenta la media, desvío estándar (DS) y coeficiente de variación (CV) de las distribuciones empíricas para las GWR entre medio-hermanos, calculadas utilizando cada uno de los cuatro métodos y los dos conjuntos de datos genómicos descriptos en la sección de Materiales y Métodos. La primera columna del cuadro corresponde a la media, DS y CV teóricos esperados para la relación entre medio-hermanos. El valor esperado para la media, condicional a la información de la genealogía y en ausencia de consanguinidad, es igual a 0,25. Por su parte, el DS teórico esperado es de 0,0373 para la relación de medio-hermanos, calculado utilizando la fórmula de Hill (1993 a). Dichos valores teóricos se tomaron de referencia al evaluar y comparar el desempeño de los cuatro métodos para estimar las GWR entre medio-hermanos. Los resultados presentados en el Cuadro 4.1 sugieren que los tres métodos que emplean información genómica y genealogía (Guo, Elsen y Han y Abney) produjeron estimaciones con valores medios muy cercanos al esperado. Los mismos se encontraron en un estrecho rango comprendido entre 0,2405 para el método de Han y Abney (2011) utilizando la base de datos DMAF_0,01 y 0,2526 para el método de Guo (1994) sobre el mismo conjunto de datos. Ahora bien, dentro de este intervalo estrecho de valores se hallaron dos tipos de diferencias significativas: i) entre las medias de los diferentes métodos empleando la misma base de datos genómica y ii) entre las medias observadas al variar la cantidad de información genómica (DMAF_0,01 *vs* DMAF_0,20) empleando el mismo método de estimación. Estas diferencias se indican con letras diferentes en el Cuadro 4.1. Contrario a lo ocurrido con los métodos que emplean información genealógica y molecular, la metodología de VanRaden (2008) produjo distribuciones empíricas para las GWR con medias de 0,2713 y 0,2685 para DMAF_0,01 y DMAF_0,20, respectivamente. Nótese que estos valores son significativamente mayores a los obtenidos empleando las otras metodologías y al esperado para la relación de medio-hermanos.

**Cuadro 4.1.** Media, desvío estándar (DS) y coeficiente de variación (CV) para la distribución empírica de las relaciones realizadas entre medio-hermanos obtenidas mediante cuatro métodos de estimación, sobre dos bases de datos genómicos editados según valores de MAF diferentes (0,01 y 0,20). La primera columna corresponde a los valores teóricos esperados.

| | **Teoría** (Hill, 1993a) | **Método de estimación** | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | **Guo (1994)** | | **Han y Abney (2011)** | | **Elsen *et al*. (2009)** | | **VanRaden (2008)** | |
| **MAF** | | **0,01** | **0,20** | **0,01** | **0,20** | **0,01** | **0,20** | **0,01** | **0,20** |
| **Media** | 0,2500 | 0,2529[a] | 0,2526[a,d] | 0,2405[b] | 0,2497[c] | 0,2516[a,e] | 0,2513[d,e] | 0,2713[f] | 0,2685[g] |
| **DS** | 0,0373 | 0,0217[h] | 0,0219[h] | 0,0322[i] | 0,0326[i] | 0,0290[i] | 0,0287[i] | 0,0927[j] | 0,0918[j] |
| **CV** | 0,1492 | 0,0858 | 0,0867 | 0,1339 | 0,1306 | 0,1152 | 0,1142 | 0,3417 | 0,3419 |

Aquellos valores con letras distintas son significativamente diferentes. DS: desvío standard. CV: coeficiente de variación.

Ahora bien, al examinar en el Cuadro 4.1 lo ocurrido en términos del DS se observa que el método propuesto por Guo (1994) fue aquel que presentó el menor valor. Al contrastar los valores de los DS para las GWR estimadas empleando cada método con respecto al valor esperado según Hill (1993 a), se observa que los métodos de Han y Abney (2011) y de Elsen *et al.* (2009) fueron los más cercanos a dicho valor teórico, sin presentar diferencias significativas entre ellos. Nótese que existieron diferencias significativas en términos de DS entre estos dos métodos y los dos restantes a valores de MAF constantes. El método de VanRaden (2008) presentó los mayores DS y más lejanos al valor teórico esperado. La misma tendencia se observó para el CV dado que depende el valor que toma el DS. El método de Han y Abney (2011), seguido por el de Elsen *et al.* (2009), presentaron los valores más cercanos al CV esperado. En cambio, el método de Guo (1994) produjo valores muy bajos de CV asociados a la baja variabilidad capturada por el método. En la Figura 4.1 se presentan diagramas de cajas y bigotes en los que pueden apreciarse las diferencias mencionadas entre los métodos, especialmente notorias a nivel de DS.

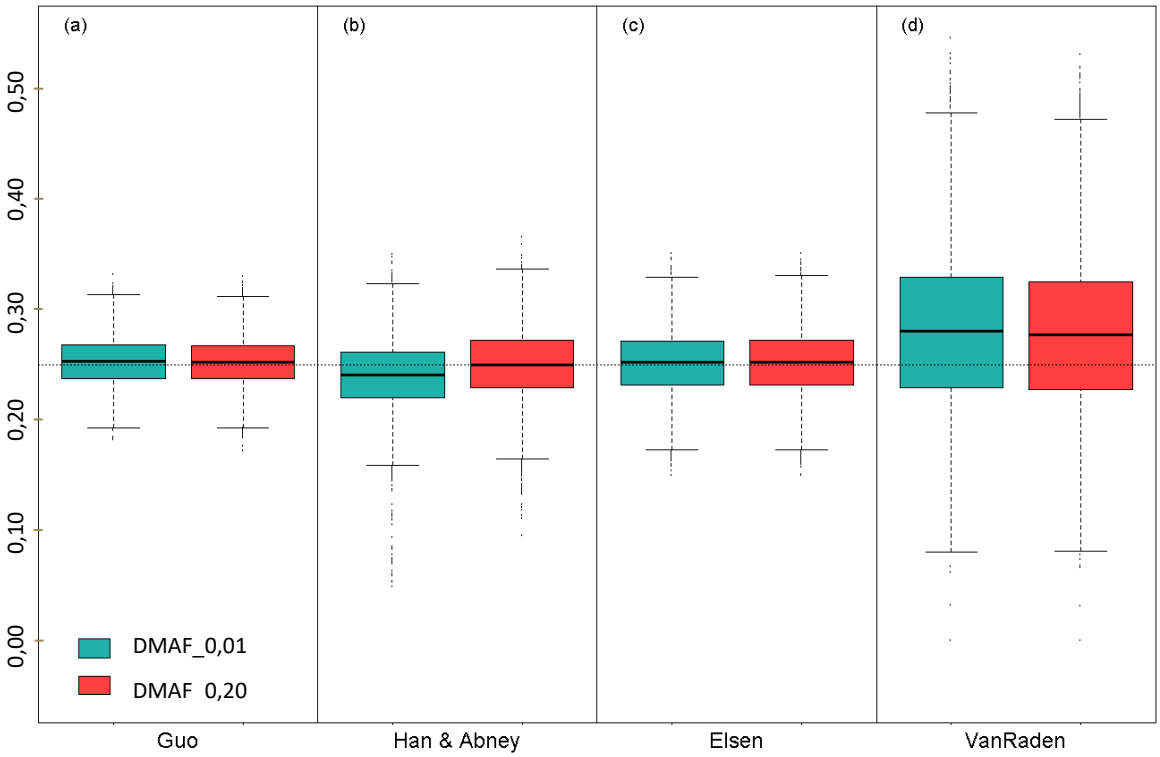



**Figura 4.1.** Diagramas de cajas y bigotes para los cuatro métodos de estimación de las relaciones de parentesco realizadas empleando ambos conjuntos de datos (DMAF_0,01 and DMAF_0,20) **a.** Método de Guo (1994), **b.** Método de Han y Abney (2011), **c.** Método de Elsen *et al.* (2009) y **d.** Método de VanRaden (2008).

En la Figura 4.2 se presentan las distribuciones empíricas obtenidas para las relaciones de parentesco estimadas por los cuatro métodos empleando ambos conjuntos de datos (DMAF_0,01 y DMAF_0,20). En la misma se incluye la distribución teórica esperada para la relación de parentesco analizada asumiendo una normal con media 0,25 que corresponde al valor esperado para medio-hermanos; la varianza corresponde a 0,0014, calculada siguiendo la expresión propuesta por Hill (1993 a). Visscher (2009) comparó los valores teóricos esperados con aquellos empíricos obtenidos en su trabajo con datos reales de humanos y mostró que las varianzas empíricas resultaron menores (aproximadamente un 16% menos) que los valores teóricos esperados por Hill (1993 a). Esta diferencia tiene origen en los supuestos que emplea Hill (1993 a) a la hora de calcular la varianza teórica, aspecto que será discutido en el próximo capítulo. En la Figura 4.2 se presenta una distribución (en color gris) con una media de 0,25 y una varianza 16% inferior a la teórica calculada con la fórmula de Hill (1993 a). Según Visscher (2009) se deberían observar curvas similares a esta al emplear datos reales para la estimación de las relaciones de parentesco realizadas. Nótese que para el caso de VanRaden (2008) la curva obtenida se encuentra desplazada hacia la derecha con una media mayor a 0,25 para ambos conjuntos de datos, tal como se presentó en el Cuadro 4.1. Además, la distribución muestra una gran cantidad de casos con valores cercanos a cero para la GWR, aspecto que impacta considerablemente en la variabilidad de la misma. Nótese además, que el método de VanRaden (2008) es el único de los cuatro que produjo valores inferiores a cero para ambos

conjuntos de datos genómicos. Todos los métodos que emplean información genómica y de genealogía (Guo, Elsen y Han y Abney) presentan distribuciones empíricas muy cercanas a la teórica (en negro) y a la referenciada por Visscher (2009) (en gris), principalmente respecto de la media. Las mayores diferencias se observaron para la varianza de la GWR, donde el método de Guo (1994) mostró el menor valor y fue significativamente distinto de los de Han y Abney (2011) y Elsen *et al.* (2009), estos dos últimos con valores más cercanos al esperado. Por su parte, el método de VanRaden (2008) presentó la mayor variabilidad entre los cuatro métodos. Nótese que estas tendencias se observaron para las estimaciones obtenidas empleando ambos conjuntos de datos (DMAF_0,01y DMAF_0,20).

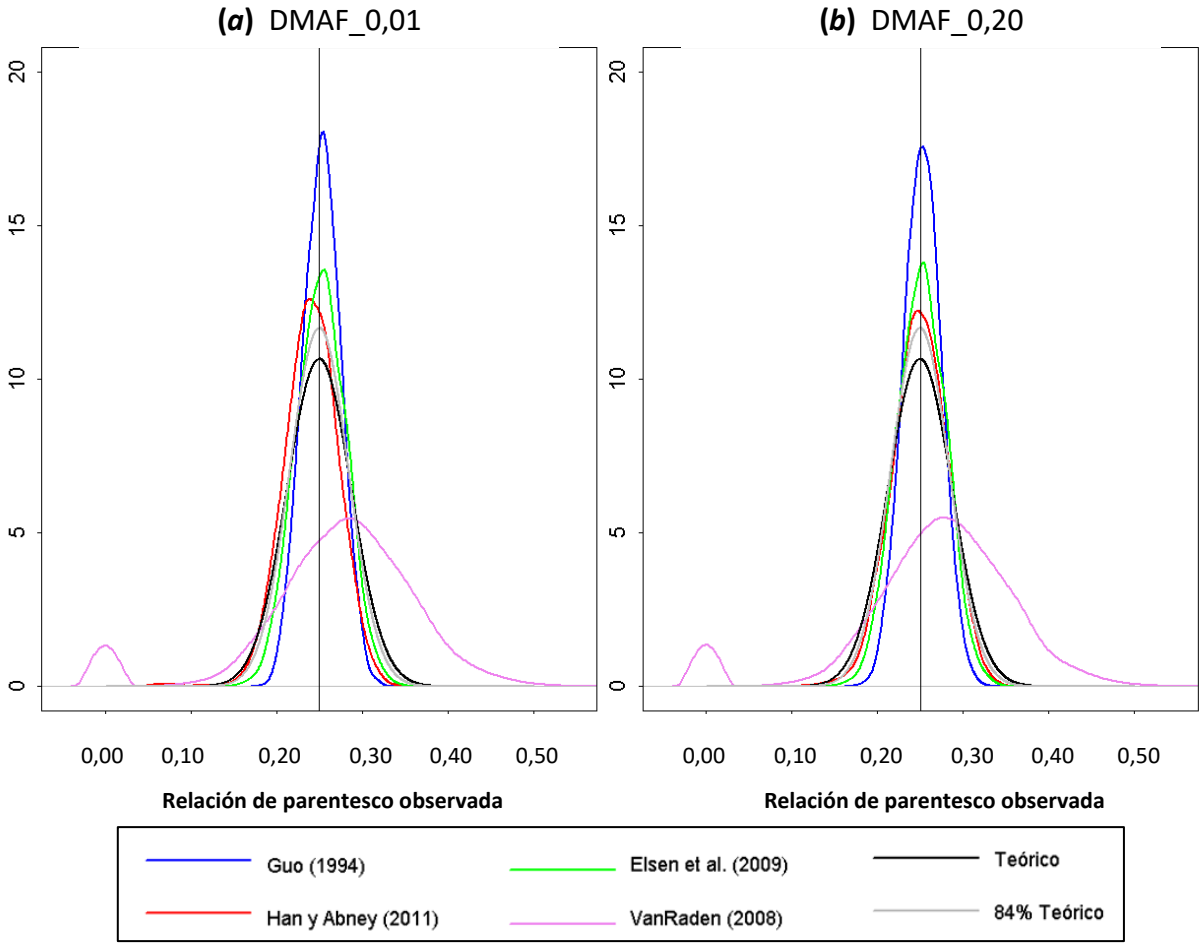



**Figura 4.2.** Distribuciones empíricas de las relaciones de parentesco estimadas empleando los métodos de Guo (1994), Han y Abney (2011), Elsen *et al.* (2009) y VanRaden (2008) con el conjunto de datos **(a)** DMAF_0,01 y **(b)** DMAF_0,20. En negro se presenta la distribución teórica esperada y en gris, la distribución teórica con menor variabilidad (un 84 % de la esperada) según fue reportada por Visscher (2006).

En la Figura 4.3 se contrasta las estimaciones de las GWR con DMAF_0,01 producidas por cada método. En este caso la comparación se realiza de a dos métodos tomando las seis combinaciones posibles de los cuatro métodos. Cada punto en las figuras representa una pareja de medio-hermanos. En una situación ideal en la que se obtienen las mismas estimaciones, independientemente del método empleado, los puntos deberían concentrarse cercanos al valor 0,25, tanto para el eje *y* como para el *x*, con cierta dispersión a lo largo de una recta con pendiente uno y ordenada al origen cero. Dicho de otro modo, debería observarse una asociación con tendencia positiva. Por ejemplo, aquella pareja que obtuvo una estimación de 0,23 con el primer método debería presentar el mismo valor al estimarlo con la segunda metodología. Nótese que el gráfico (f) es el que más se acerca a esta situación. En el mismo se presentan las estimaciones obtenidas con los métodos de VanRaden (2008) y Han y Abney (2011). Tanto en la figura (a) como en la (c) aumenta la dispersión de los puntos a lo largo del eje *y* asociado al método de Han y Abney (2011). Nótese que en ambos casos existe una nube de puntos menos densa por debajo del valor $y = 0,15$ que corresponde a un conjunto de pares de medio-hermanos que presentaron estimaciones de GWR muy bajas con el método de Han y Abney (2011). Los mismos pares tuvieron estimaciones entre 0,20 y 0,30 con los procedimientos de Guo (1994) y de Elsen *et al.* (2009). Por otro lado, al analizar los gráficos (d) y (e) y (f) es notorio el aumento en la variabilidad generado por el método de VanRaden (2008) con respecto a las otras metodologías. Nótese que, en los tres casos, la nube de puntos presenta mayor dispersión siguiendo el eje *y* asociado con el método de VanRaden (2008), en comparación a lo que ocurre en el eje *x*. Se observa, además, una línea de puntos en la parte inferior de los tres gráficos (d, e y f) para $y = 0$, lo que pone de manifiesto que el método de VanRaden (2008) produjo valores cero para GWR cuyas estimaciones estuvieron entre 0,05 y 0,35 con los métodos que utilizan información de genealogía y de marcadores (Guo, Elsen y Han y Abney). Por último, el grafico (f) presenta un patrón distinto a los otros cinco en la Figura 4.3. Nótese que en este caso se observa una leve asociación positiva entre las estimaciones producidas por VanRaden (2008) y Han y Abney (2011). Los mismos gráficos se obtuvieron para DMAF_0,20 y se presentan en la Figura 4.4. Nótese que las tendencias observadas con DMAF_0,01 se mantuvieron con DMAF_0,20.

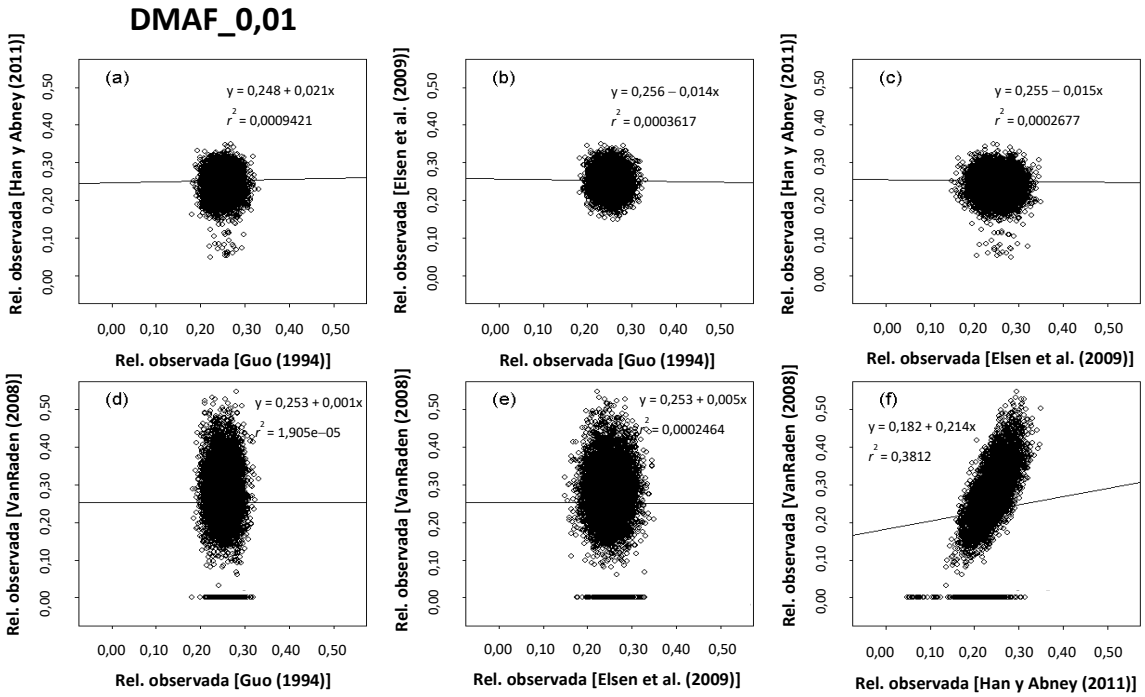



**Figura 4.3.** Comparación de las relaciones de parentesco estimadas empleando DMAF_0,01 entre los diferentes métodos: **(a)** Guo (1994) y Han y Abney (2011); **(b)** Guo (1994) y Elsen *et al.* (2009); **(c)** Elsen *et al.* (2009) y Han y Abney (2011); **(d)** Guo (1994) y VanRaden (2008); **(e)** Elsen *et al.* (2009) y VanRaden (2008); **(f)** Han y Abney (2011) y VanRaden (2008).

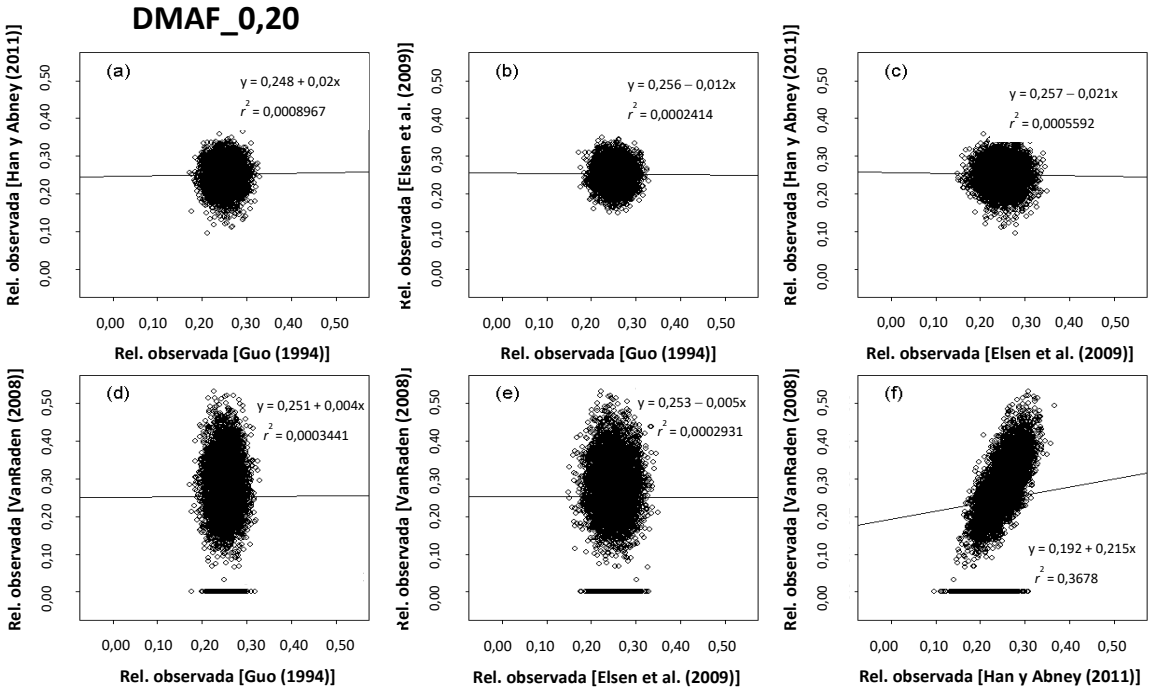



**Figura 4.4.** Comparación de las relaciones de parentesco estimadas empleando DMAF_0,20 entre los diferentes métodos: **(a)** Guo (1994) y Han y Abney (2011); **(b)** Guo (1994) y Elsen *et al.* (2009); **(c)** Elsen *et al.* (2009) y Han y Abney (2011); **(d)** Guo (1994) y VanRaden (2008); **(e)** Elsen *et al.* (2009) y VanRaden (2008); **(f)** Han y Abney (2011) y VanRaden (2008).

## 4.2. Desempeño de los métodos de estimación frente a variaciones en la cantidad de información genómica disponible

Para evaluar la sensibilidad de los distintos métodos a cambios en la cantidad de información de marcadores genómicos es necesario analizar lo ocurrido con las estimaciones al emplear DMAF_0,01 y DMAF_0,20. Tal como reporta el Cuadro 4.1, sólo los métodos de Han y Abney (2011) y VanRaden (2008) presentaron diferencias significativas en las medias al cambiar el conjunto de datos. Para el caso de la varianza no se encontraron diferencias significativas entre ambos conjuntos de datos. Los dos métodos restantes no presentaron diferencias significativas en las medias ni en el DS de las GWR cuando se calcularon con DMAF_0,01 o con DMAF_0,20. Nótese que para el DS se encontraron diferencias significativas sólo entre métodos, pero no entre las estimaciones producidas con DMAF_0,01 y DMAF_0,20 para cada método. Entonces, la cantidad de información genómica empleada para estimar GWR no afectó el DS de las distribuciones empíricas obtenidas con cada método. Entre aquellos procedimientos que emplean información de marcadores moleculares y de genealogía, el de Han y Abney (2011) tuvo un comportamiento fuera de lo esperado dado que estimó GWR muy bajas a partir de la base DMAF_0,01 para un número importante de pares de individuos que no parecían encontrarse pobremente emparentados cuando se observaban las estimaciones obtenidas con los otros métodos (Figura 4.3 a y c). Nótese además los valores extremadamente bajos que alcanzan los bigotes en la Figura 4.1 (b). Para notar la magnitud de dicha diferencia tómese por ejemplo un valor de GWR igual a 0,15. Con el método de Han y Abney (2011) se estimaron GWR para 36 parejas de medio-hermanos por debajo de 0,15 con DMAF_0,01 y 25 con DMAF_0,20. Por el contrario, el método de Elsen *et al.* (2009) sólo produjo dos estimaciones inferiores a 0,15 con DMAF_0,01 y una con DMAF_0,20. La metodología de Guo (1994) no produjo ningún valor inferior a 0,15. Fuera ya de los métodos que utilizan información de marcadores y genealogía, el método de VanRaden (2008) generó estimaciones inferiores a 0,15 para más de 470 parejas, lo que se refleja directamente en la magnitud del SD (Figura 4.1 y Cuadro 4.1).

En la Figura 4.5 se presenta la regresión de las estimaciones de GWR producidas empleando DMAF_0,01 sobre aquellas obtenidas con DMAF_0,20 para cada método. De este modo es posible examinar con mayor grado de detalle la sensibilidad de cada una de las metodologías para captar la relación realizada empleando diferentes cantidades de información genómica. Cuando la cantidad de información molecular disponible no afecta la estimación de GWR, la correlación entre las estimaciones obtenidas mediante ambos conjuntos de datos es similar al coeficiente de regresión. La Figura 4.5 muestra un acuerdo notorio entre las estimaciones para ambos conjuntos de datos genómicos al emplear las metodologías de Elsen *et al.* (2009) y VanRaden (2008), con una correlación de aproximadamente 0,99 y un coeficiente de regresión muy cercano a la unidad. Por su parte, la correlación para el método de Han y Abney (2011) fue más baja (0,93 aproximadamente) así como también su coeficiente de regresión, debido principalmente al desempeño del mismo al estimar las GWR de aquellos pares de medio-hermanos cuya relación realizada es inferior al valor esperado. De este modo se pone de manifiesto la sensibilidad de éste método a cambios en la cantidad de información molecular al trabajar con estos casos puntuales. De hecho, aquellas parejas de medio-hermanos con una GWR estimada en 0,15 al emplear DMAF_0,20, obtuvieron estimaciones en torno a 0,05 al usar el otro conjunto de

datos genómicos. Finalmente, el método de Guo (1994) fue aquel que presentó los valores más bajos de correlación (0,84) y de coeficiente de regresión (0,905), mostrando de este modo una mayor sensibilidad a los cambios en la cantidad de información genómica disponible, en comparación con las otras tres metodologías.

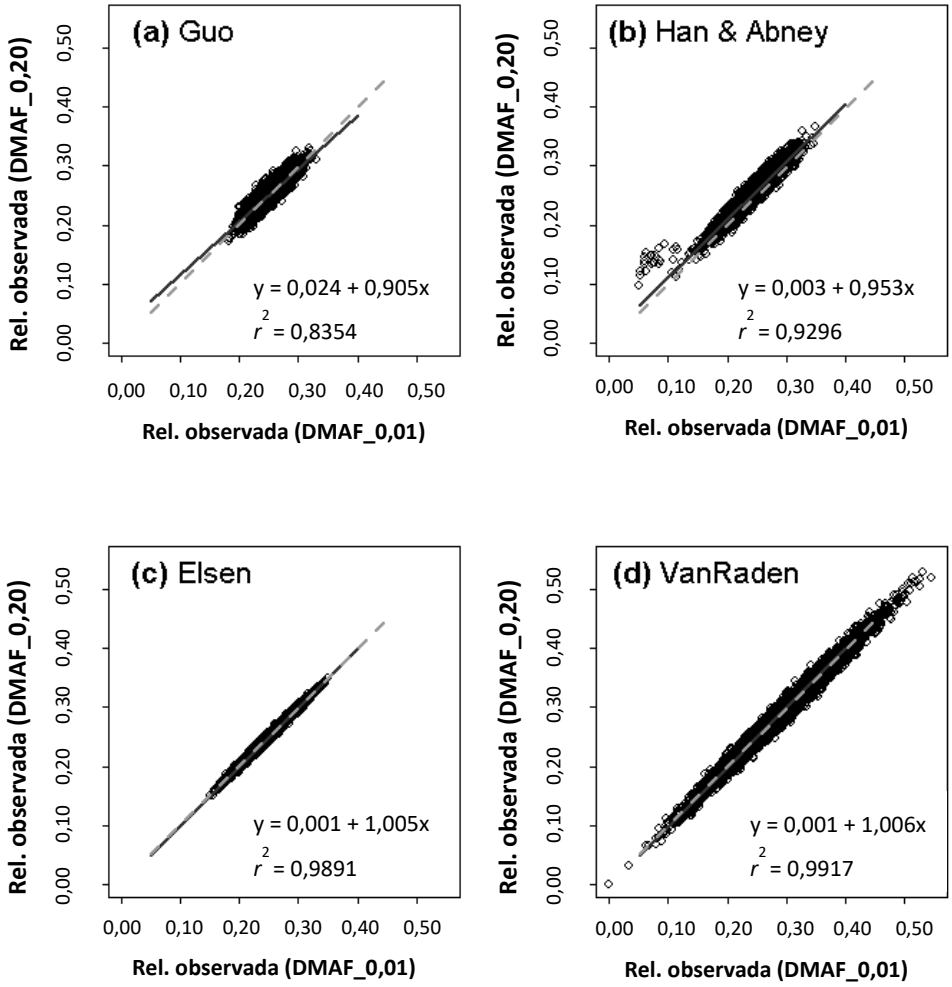



**Figura 4.5.** Regresión de las estimaciones de GWR producidas empleando DMAF_0,20 en aquellas obtenidas al usar DMAF_0,01 para cada uno de los cuatro métodos: **a.** Guo (1994), **b.** Han y Abney (2011), **c.** Elsen *et al.* (2009) y **d.** VanRaden (2008). La recta gris de guiones representa el caso ideal en el cual las estimaciones de las relaciones de parentesco realizadas no varían según el conjunto de datos utilizado.

A continuación se presenta la Figura 4.6 donde se observa el efecto de emplear dos bases de datos genómicas diferentes (DMAF_0,01 y DMAF_0,20) sobre las distribuciones empíricas de las GWR estimadas por cada método. Puntualmente para el caso de Guo

(1994) y Elsen *et al.* (2009) (recuadro a y c, respectivamente) no se observan diferencias en términos de la media o la varianza de las distribuciones empíricas al estimar las GWR usando DMAF_0,01 y DMAF_0,20, tal como se reporta en el Cuadro 4.1. Por su parte, en el caso de Han y Abney (2011) (recuadro b) se advierte un leve desplazamiento hacia la izquierda de la curva para DMAF_0,01. Esto está asociado con aquellos pares de medio-hermanos con GWR estimadas con valores muy bajos (inferiores a 0,15). No se observaron cambios en la varianza de las estimaciones al modificar la cantidad de información genómica empleada con este método. Finalmente, en el caso de las GWR estimadas con el método de VanRaden (2008) (recuadro d), la curva para DMAF_0,01 se encuentra desplazada levemente hacia la derecha con respecto a DMAF_0,20. Esto genera diferencias significativas en las medias pero no a nivel de varianza, tal como se observa en el Cuadro 4.1.

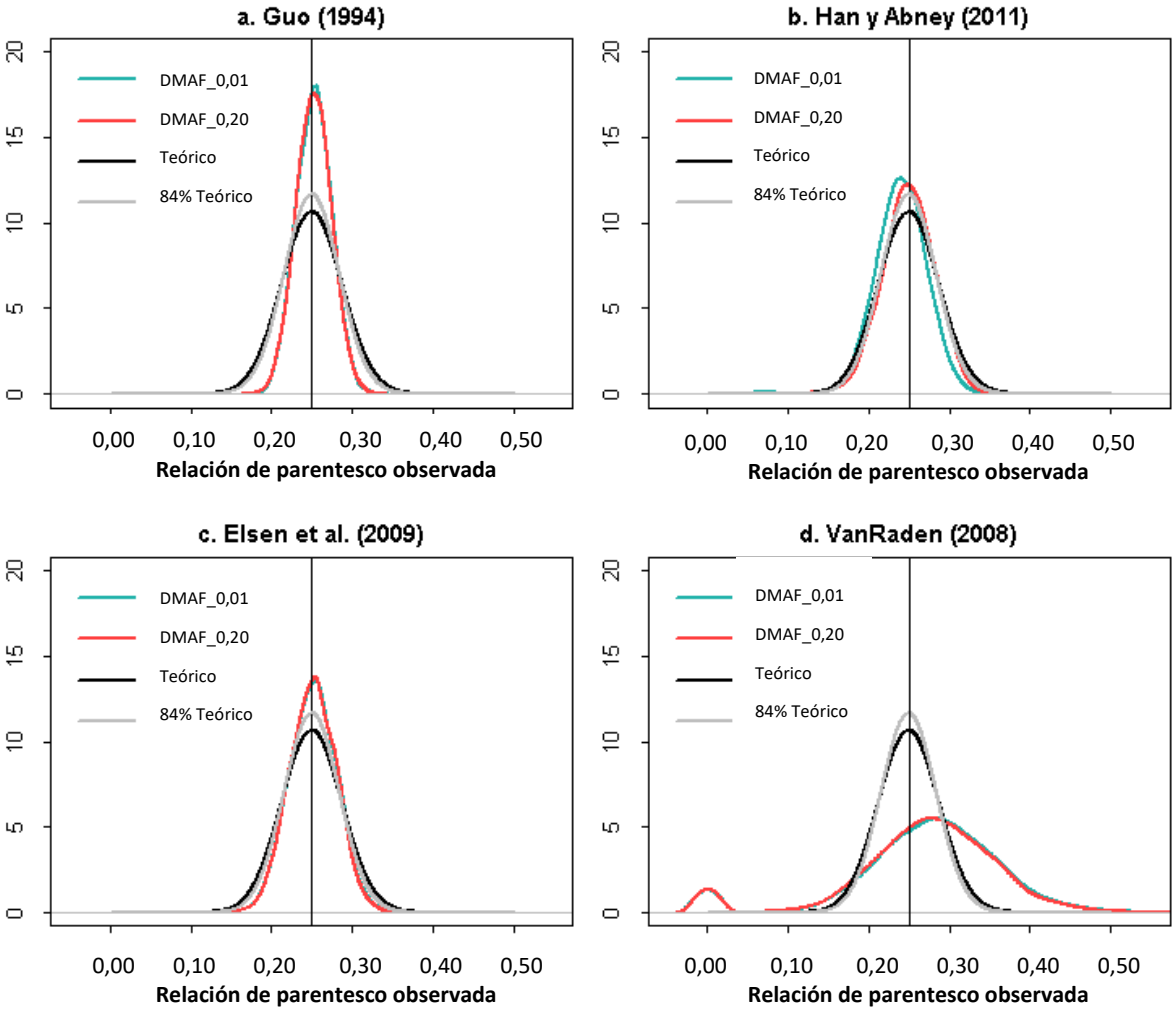



**Figura 4.6.** Distribuciones empíricas de las GWRs estimadas empleando los conjuntos de datos DMAF_0,01 y DMAF_0,20 para los métodos de **a.** Guo (1994), **b.** Han y Abney (2011), **c.** Elsen *et al.* (2009) y **d.** VanRaden (2008). En negro se presenta la distribución teórica esperada y en gris, la distribución teórica con menor variabilidad (un 84 % de la esperada) según lo reportado por Visscher (2006).

# Capítulo 5. *Discusión*

# Capítulo 5

# Discusión

La disponibilidad de datos genómicos permite estimar las GWR entre individuos porque hace posible detectar qué proporción real del genoma comparten. En la actualidad existe una amplia variedad de métodos para estimar GWR. En este trabajo se evaluaron cuatro metodologías diferentes para estimar GWR entre medio-hermanos de la generación $F_2$ de una población experimental de cerdos utilizando dos conjuntos de datos genómicos con diferentes cantidades de información (DMAF_0,01 y DMAF_0,20). Las estimaciones resultantes fueron evaluadas y comparadas en términos de su distribución empírica. Se observó que los tres métodos basados en un enfoque IBD, que emplean información de marcadores moleculares y de genealogía, generaron estimaciones cuyo promedio observado fue muy cercano al valor esperado para la relación de parentesco analizada. Por el contrario, las estimaciones obtenidas con el método basado en IBS, que solo emplea información molecular (VanRaden, 2008), presentaron una media mayor y distinta del valor esperado. Si bien el método de VanRaden (2008) permite estimar la proporción de genoma compartido IBD (Toro *et al.*, 2011), no considera que el ADN se transmite por segmentos y no por genes o puntos independientes (Thompson, 2013). De hecho, al permutar los marcadores entre sí las estimaciones producidas por este método no varían. Esto se debe a que el método no considera el ligamiento y el desequilibrio gamético. Además, el desempeño del método de VanRaden (2008) fue afectado por la estructura de la población experimental de cerdos. Esto se debe, principalmente, a cómo fue el muestreo de las frecuencias alélicas en dicha población y a la sensibilidad del método en relación a las mismas.

En términos de la variabilidad de las GWR entre medio-hermanos, los métodos de Han y Abney (2011) y Elsen *et al.* (2009) estimaron relaciones con DS más cercanos al teórico esperado en comparación con los DS obtenidos empleando los métodos de Guo (1994) y VanRaden (2008). De hecho, el método de Guo (1994) sólo alcanzó un 34 % del valor teórico esperado para el tipo de relación de parentesco analizada. Esto se debe principalmente a que esta metodología asigna una PIBD igual a 0,5 a aquellos segmentos genómicos en los que uno o ambos marcadores moleculares que lo delimitan, no son completamente informativos o poseen genotipos faltantes. Si bien este modo de proceder garantiza que la media de la distribución se mantenga en torno al valor esperado para medio-hermanos, también concentra las estimaciones en torno a esta cifra. En consecuencia, se reduce la variabilidad y se pierde sensibilidad para detectar relaciones extremas entre medio-hermanos. Finalmente, el método de VanRaden (2008) mostró varianzas significativamente mayores a todas las otras metodologías y a la teórica esperada. Esto se debe a que en este trabajo se emplearon las frecuencias alélicas calculadas en la generación $F_0$ para construir la matriz $\boldsymbol{G}$, tal como lo señala VanRaden (2008). Ahora bien, nótese que las frecuencias alélicas se estimaron a partir de un conjunto reducido de animales provenientes de dos razas diferentes (15 hembras Pietrain y cuatro machos Duroc). Al comparar las frecuencias alélicas de ambas razas se encontraron diferencias

significativas para varios loci. En consecuencia, el exceso de variabilidad en las frecuencias alélicas de la generación inicial pudo contribuir al aumento en la varianza de las GWR de individuos de la generación $F_2$. Nótese, además, que dichas frecuencias se computaron empleando la información de un número muy reducido de individuos de la generación base ($F_0$). Idealmente, las frecuencias alélicas deberían calcularse tomando a todos los individuos de cada una de las poblaciones (razas puras) de las cuales provinieron esos 19 animales base. Ahora bien, la escasa información disponible (sólo 19 animales base) puede emplearse para obtener estimaciones sencillas de las frecuencias alélicas, pero a costa de sesgo (VanRaden, 2008). Como consecuencia, el método de VanRaden (2008) fue el más afectado de los cuatro usados para estimar GWR por dos razones: i) fue desarrollado para una población única (no para cruzas), y ii) es una metodología sensible a las frecuencias alélicas empleadas para llevar a cabo los cálculos, tal como discute VanRaden (2008). Esto se debe a que el método emplea solamente información molecular sin considerar la genealogía en el cálculo. Wakeley *et al.* (2012) observó recientemente que el pedigrí de un organismo diploide debe ser tratado como parámetro de la distribución del proceso de herencia porque, si bien no contiene toda la información sobre la trasmisión del genoma entre generaciones, restringe el espacio paramétrico en donde son informativos los marcadores moleculares. Entonces, la igualdad de alelos en ciertos marcadores moleculares entre individuos (o *identidad en estado*) por sí sola no es evidencia de herencia compartida y genera error respecto de la información transmitida.

Los tres métodos que emplean la información de marcadores moleculares y genealogía para estimar GWR reportaron varianzas empíricas por debajo del valor teórico esperado: ninguno alcanzó más del 76% de dicho valor. Visscher (2009) observó una varianza empírica 16% inferior al valor teórico, en un conjunto de 4401 pares de hermanos enteros humanos. Estas diferencias entre los valores empíricos y teóricos tienen su origen en los supuestos subyacentes a la teoría, particularmente en el supuesto de ausencia de interferencia y empleo de la función de Haldane (1919) para modelar la localización para las recombinaciones a lo largo de los cromosomas. Dicha función emplea una distribución uniforme de las posiciones de los eventos de recombinación, pero existen otros enfoques en los que pueden modelarse empleando una distribución localizada. Esta última describe mejor el proceso genético subyacente ya que se demostró en diversas especies que las localizaciones de las recombinaciones no se distribuyen uniformemente a lo largo del genoma. Contrario a esto, existen regiones cromosómicas denominadas *hot spots*, en las cuales los eventos de recombinación ocurren con mayor frecuencia que en otros lugares del genoma (Ma *et al.*, 2015). Tortereau *et al.* (2012) encontraron, en poblaciones de cerdos, que la distribución de las tasas de recombinación no es uniforme a lo largo de los cromosomas. Las mayores tasas se encuentran concentradas en la región final de los mismos. Es decir que en los extremos cromosómicos la probabilidad de observar un evento de recombinación es mayor que en otros puntos. Esta incidencia no aleatoria de las recombinaciones complica la detección de los segmentos de IBD entre individuos porque requiere un número de marcadores 12 veces mayor al número de segmentos IBD presentes para detectar el 90% de los mismos (MacLeod *et al.*, 2005). Por otra parte, el enfoque de una distribución localizada de las recombinaciones conlleva una mayor dependencia entre loci, aumentando la varianza de la PIBD global. Por ejemplo, Suarez *et al.* (1979) obtuvieron un DS de 0,055 para PIBD entre hermanos enteros en humanos al asumir una distribución localizada de las recombinaciones, mientras que al emplear la función de

Haldane (1919), con los mismos datos, se obtuvo un DS de 0,040 (Risch y Lange, 1979). De similar modo, los resultados del presente trabajo indican que los métodos de Guo (1994) y Elsen *et al.* (2009), que utilizan para el cálculo la función de recombinación de Haldane, presentaron distribuciones empíricas de GWR estimadas con menor variabilidad, i.e. menor DS.

Los resultados obtenidos en este trabajo son comparables en orden de magnitud a aquellos informados por Visscher (2009) y Guo (1996) para medio-hermanos en humanos. Al inspeccionar la fórmula de la varianza para la relación estimada entre medio-hermanos según Hill (1993 a), es de esperar diferencias en la varianza de GWR para individuos de dos especies distintas. Esto se debe a que la varianza depende de mecanismos meióticos y cromosómicos específicos de cada especie, tal como se detalla más adelante en el Cuadro 5.1. En especies con menos cromosomas y un número menor de eventos de recombinación, como es el caso del cerdo con respecto al humano (Tortereau *et al.*, 2012), se observan valores mayores para la varianza de la GWR (Rasmuson, 1993). En el Cuadro 4.1 se observa que el DS empírico de las GWR estimadas con los métodos de Elsen *et al.* (2009) y Han y Abney (2011) fueron entre 2,87% a 3,26 %. Estos valores son comparativamente mayores a aquellos reportados para los datos en humanos que estuvieron entre 2,50% a 2,64%, pero son similares en orden de magnitud. Además, como observó Visscher (2009), la varianza observada en las estimaciones en cerdos es relativamente pequeña comparada con el promedio, reflejándose en valores bajos del coeficiente de variación. Los CV se presentan en el Cuadro 4.1 y oscilan entre 9 al 13 %, valores comparables con aquellos obtenidos por Visscher (2009) para medio-hermanos en humanos.

**Cuadro 5.1.** Diferencias en la estructura del genoma y en los procesos genómicos entre cerdos y humanos.

| | **Tasa de recombinación promedio** | **Largo total del genoma (promedio)** | **Cromosomas autosómicos** |
|---|---|---|---|
| **Humanos** (Kong *et al.*, 2002) | 1,13 cM/Mb | 3190,77 Mb | 22 |
| **Cerdos** (Tortereau *et al.*, 2012) | 0,76 cM/Mb | 2334,00 Mb | 18 |

**Mb**: Megabase; **cM**: centiMorgan

Por otra parte, al comparar las GWRs estimadas con dos bases de datos genómicos distintas (Figura 4.5) es necesario interpretar los resultados en función del modo en que funciona cada una de las metodologías. Aquella propuesta por Guo (1994) mostró mayor efecto sobre las estimaciones obtenidas a partir del cambio en la información genómica.

Esto se debe, en gran medida, a las características específicas del método. Esta metodología evalúa los marcadores de a pares para estimar la PIBD por segmentos, pero sin considerar el LD. Según los alelos de los marcadores que delimitan cada segmento cromosómico, el fragmento analizado corresponde a alguno de todos los casos posibles presentados en el Cuadro 2.1. Nótese que al pasar de un MAF 0,01 a un valor de 0,20 en la edición de la base de datos, los marcadores remanentes resultarán, en promedio, más informativos. Dicho de otro modo, al aumentar el MAF disminuye la probabilidad que se presenten casos tales como el 5 (Cuadro 2.1), donde un marcador es no informativo y la incertidumbre es alta. En consecuencia, al bajar la incidencia de dichos casos disminuyen los niveles de incertidumbre y aumenta la capacidad del método para captar la verdadera PIBD compartida entre medio-hermanos. Por el contrario, tanto el método de Han y Abney (2011) como el de Elsen *et al.* (2009) consideran el LD y ligamiento físico en sus cálculos. Concretamente el método de Elsen *et al.* (2009) realiza una exploración previa para determinar cuáles son los bloques de ligamiento y toma un subconjunto de marcadores en cada uno de ellos con la información mínima y suficiente para llevar a cabo los cálculos. Este modo de funcionar justifica la estabilidad en las predicciones al variar la MAF. El mismo método elimina aquellos marcadores no informativos para cada bloque de ligamiento dado que no aportan información adicional útil a la estimación y generan ineficiencias al momento del cómputo. Algo similar ocurre con el método de Han y Abney (2011) ya que para llevar a cabo las estimaciones considera conjuntamente la información de aquellos marcadores no independientes. A tal efecto, considera un número fijo de marcadores en LD cuyo valor por default es 10 loci. Esto permite emplear más información (que conlleva una disminución de la incertidumbre) para calcular la PIBD en cada segmento. Ahora bien, el método de Han y Abney (2011) se diferencia del de Elsen *et al.* (2009) porque no realiza un paso previo para determinar cuáles son los bloques de ligamiento. Simplemente toma en cuenta grupos de marcadores que presentan una elevada correlación entre ellos (10, por default), con lo cual lo observado en un locus en particular es condicional a lo ocurrido con los otros nueve loci altamente correlacionados con ese locus. De todos modos, esto no asegura que se utilizando toda la información de un bloque de ligamiento. La estrategia permite disminuir la incertidumbre en casos donde al menos uno de los marcadores en LD es completamente informativo, pero mantiene cierto nivel de incertidumbre en aquellos loci altamente correlacionados y no informativos. Este es el motivo por el cual se observa cierto efecto sobre la media de las estimaciones al pasar de DMAF_0,01 a DMAF_0,20. Se entiende que en este último caso es mayor la proporción de marcadores informativos y, por lo tanto, menor la incertidumbre asociada. Finalmente en el caso de VanRaden (2008), tal como se mencionó anteriormente, no se considera la genealogía y los marcadores moleculares se asumen independientes, consecuentemente es de esperar que las estimaciones producidas por el método se vean afectadas por la cantidad e informatividad de los datos genómicos empleados, tal como ocurrió en este trabajo.

Podemos resumir esta tesis señalando que la investigación empleó datos de una población experimental de cerdos con una genealogía relativamente sencilla y el análisis se enfocó específicamente en la relación de medio-hermanos tomándola como referencia para comparar los distintos métodos evaluados. Sin embargo, las poblaciones reales que surgen de las evaluaciones genéticas son estructuralmente más complejas, muestran niveles de consanguinidad importantes y presentan un sinfín de relaciones de parentesco. Esta consideración es importante al momento de contrastar el desempeño de los diferentes

métodos. Concretamente, el método de Guo (1994) no es útil al momento de trabajar con genealogías complejas dado que fue desarrollado para estimar las GWR para medio-hermanos: sólo sirve para esos casos y no puede emplearse para estimar otras GWR tales como primos hermanos, abuelo-nieto, etc. Si bien existe la posibilidad de extender fácilmente el método para evaluar la GWR entre hermanos enteros, no es posible hacerlo para otras más complejas, limitando así su aplicabilidad. De manera análoga, el método de Elsen *et al.* (2009) fue desarrollado para familias de medio-hermanos o para una combinación de medio-hermanos y hermanos enteros. Por su parte, el método propuesto por Han y Abney (2011) permite obtener estimaciones de las GWR para todas las relaciones posibles en una genealogía más compleja sin necesidad de modificaciones. No obstante, posee ciertas limitaciones al trabajar con grandes cantidades de animales porque emplea un algoritmo HMM. Estos modelos son eficientes para estimar PIBD en pares de individuos, siempre y cuando la genealogía contenga unas pocas centenas de individuos (Lander y Green, 1987). Para genealogías más extensas y complejas, los HMM proveen una aproximación a las GWR. De todos modos, en estos casos más complicados y con gran cantidad de marcadores moleculares, el cálculo de GWR mediante HMM se vuelve extremadamente demandante en tiempo y capacidad de cómputo, tornándose rápidamente inviable (Thompson, 2000). Por su parte, el método de VanRaden (2008) permite estimar GWR para todos los pares de individuos de una genealogía, por más que la misma sea extensa y compleja, pero con la menor precisión por no tomar en cuenta la información de pedigree ni el LD.

El desarrollo de este trabajo permitió conocer en profundidad el desempeño de cada uno de los métodos propuestos y fue de gran utilidad para conocer los diferentes enfoques a la hora de calcular las relaciones de parentesco realizadas entre individuos. Si bien el análisis se limitó a la relación entre medio-hermanos, fue suficiente para detectar variaciones en las estimaciones para una de las relaciones de parentesco más frecuentes de las poblaciones de animales domésticos. Partiendo de los resultados obtenidos en este trabajo surgen nuevos interrogantes. Entre ellos se encuentra analizar el desempeño de cada método pero considerando una mayor diversidad de relaciones de parentesco, en especial aquellas más lejanas (un número mayor de meiosis entre ambos individuos) donde la variación es mucho mayor. A tal fin sería de gran utilidad trabajar con datos simulados con el objetivo de conocer los valores reales de cada relación realizada (proporción de genoma IBD real compartido entre los individuos bajo análisis), para luego contrastarlos con los valores estimados por cada método. Cabe destacar que dicho análisis sólo sería posible de llevar a cabo con aquellos métodos que permiten estimar varios tipos de relaciones como ser el de Han y Abney (2011) y el de VanRaden (2008). Por otro lado, cabe destacar que este trabajo fue concebido y desarrollado en el marco de un proyecto de investigación de mayor alcance cuyo objetivo es el de estudiar diferentes metodologías de estimación de PIBD y proponer nuevas que permitan obtener estimaciones precisas empleando toda la información disponible eficientemente. Es así que los resultados de este trabajo contribuyen de modo directo a conocer y comprender el funcionamiento de diferentes métodos de estimación y sirven como base para proponer nuevas metodologías que cuenten con las ventajas de aquellas analizadas en esta tesis y con mejoras basadas en los problemas detectados en este trabajo. Consecuentemente, tanto este trabajo como otros del grupo de investigación sientan las bases sobre las que se pretende seguir trabajando con el objetivo de desarrollar metodologías alternativas para la estimación de PIBD.

# Capítulo 6. *Conclusiones*

# Capítulo 6

# Conclusiones

En el presente trabajo fue posible obtener estimaciones de las relaciones de parentesco realizadas, o GWR, entre medio-hermanos de una población experimental de cerdos empleando diferentes métodos. Las diferencias más generales y notorias se dieron entre aquellas metodologías que emplean conjuntamente la genealogía y la información genómica siguiendo un enfoque IBD, respecto del método que emplea solo la información molecular (IBS). Las primeras fueron las que produjeron mejores estimaciones de las GWR en términos de sus distribuciones empíricas con respecto a la teórica esperada. La misma apreciación puede hacerse para aquellos métodos que toman en cuenta el LD o ligamiento físico para el cálculo (metodologías de Han y Abney, 2011 y Elsen *et al.*, 2009). Estos procedimientos condujeron a estimaciones de GWR mejores en términos de su distribución empírica, al compararlas con la teórica esperada. La causa de este comportamiento radica en que al considerar la información de marcadores asociados se puede modelar fehacientemente el proceso genético de herencia por segmentos, sin necesidad de simplificarlo asumiendo independencia total entre los SNPs.

Los resultados obtenidos y discutidos en los capítulos precedentes sugieren que el método de Elsen *et al.* (2009) fue el que mostró el mejor desempeño a la hora de estimar las GWR entre medio-hermanos. Nótese que el mismo emplea un enfoque IBD en los cálculos combinando la información genómica con la de genealogía y considera, además, el ligamiento entre marcadores. Dado el modo en que toma en cuenta los bloques de ligamento, esta metodología puede emplear diferentes cantidades de información genómica sin afectar la calidad de las estimaciones. Tanto la media como la varianza de la distribución empírica de las GWR estimadas por el método de Elsen *et al.* (2009) fueron las más cercanas a los valores teóricos esperados y concuerdan con aquellos reportados en la literatura.

# BIBLIOGRAFÍA


Abney, M. 2008. Identity-by-descent estimation and mapping of qualitative traits in large, complex pedigrees. *Genetics*. 179: 1577–1590.

Abney, M. 2009. A graphical algorithm for fast computation of identity coefficients and generalized kinship coefficients. *Bioinformatics*. 25: 1561– 1563.

Aguilar I., Misztal I., Legarra A. y Tsuruta S. 2011. Efficient computation of the genomic relationship matrix and other matrices used in single-step evaluation. *J. Anim. Breed. Genet.* 128: 422–428.

Albrechtsen, A., Sand Korneliussen, T., Moltke, I., van Overseem Hansen, T., Nielsen, F. C. y Nielsen, R. 2009. Relatedness mapping and tracts of relatedness for genome-wide data in the presence of linkage disequilibrium. *Genet. Epidemiol.* 33: 266–274.

Anderson, A. D. y Weir, B. S. 2007. A maximum-likelihood method for the estimation of pairwise relatedness in structured populations. *Genetics.* 176: 421–440.

Bickeböller, H. y Thompsom, E. A. 1996. Distribution of genome shared IBD by half-sibs: approximation by the poisson clumping heuristic. *Theoret. Pop. Biol*. 50: 66–90.

Browning, S.R., Browning, B.L. 2007. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *Am. J. Hum. Genet.* 81: 1084–1097.

Browning, B. L. y Browning, S. R. 2009. A unified approach to genotype imputation and haplotype phase inference for large data sets of trios and unrelated individuals. *Am. J. Hum. Genet.* 84: 210–223.

Browning, S. R. y Browning, B. L. 2010. High-resolution detection of identity by descent in unrelated individuals. *Am. J. Hum. Genet.* 86: 526–539.

Darlington, C. D. 1939. *The evolution of genetic systems.* Cambridge University Press. Cambridge, United Kingdom.

Donnelly, K. 1983. The probability that related individuals share some section of genome identical by descent. *Theoret. Pop. Biol*. 23: 34–64.

Durbin, R., Eddy, S., Krogh, A. y Mitchison, G. 1998. *Biological sequence analysis: Probabilistic models of proteins and nucleic acids.* Cambridge University Press. Cambridge, United Kingdom.

Edwards, D. B., Ernst, C. W., Tempelman, R. J., Rosa, G. J. M., Raney, N. E., Hoge, M. D. y Bates, R. O. 2008. Quantitative trait loci mapping in an F2 Duroc x Pietrain resource population: I. Growth traits. *J. Anim. Sci.* 86: 241–253.

Elsen, J. M., Mangin, B., Goffinet, B., Boichard, D. y Le Roy, P. 1999. Alternative models for QTL detection in livestock - I General introduction. *Genet. Sel. Evol.* 31: 213-224.

Elsen, J. M., Filangi, O., Gilbert, H., Le Roy, P. y Moreno, C. 2009. A fast algorithm for estimating transmission probabilities in QTL detection designs with dense maps. *Genet. Sel. Evol.* 41: 50.

Favier, A., Elsen, J. M., de Givry, S. y Legarra, A. 2010. Exact haplotype reconstruction in half-sibs families with dense marker maps. En *Proceedings of World Congress on Genetics Applied to Livestock Production.* 1 – 6 de agosto de 2010, Leipzig, Alemania.

Fernandez, S. A., Fernando, R. L., Guldbrandtsen, B., Totir, L. R. y Carriquiry, A. L. 2001. Sampling genotypes in large pedigrees with loops. *Genet. Select. Evol.* 33: 337–367.

Filangi, O., Elsen, J. M., Gilbert, H., Legarra, A., Le Roy, P. y Moreno, C. 2010. QTLMap: a software for QTL detection in outbred populations. En *Proceedings of World Congress on Genetics Applied to Livestock Production*. 1 – 6 de agosto de 2010, Leipzig, Alemania.

Fishelson, M. y Geiger, D. 2002. Exact genetic linkage computations for general pedigrees. *Bioinformatics.* 18: S189–S198.

Gagnon, A., Beise, J. y Vaupel, J. W. 2005. Genome-wide identity-by-descent sharing among CEPH siblings. *Genet. Epidemiol*. 29: 215–224.

García-Cortés, L. A., Legarra, A., Chevalet, C. y Toro, M. A. 2013. Variance and covariance of actual relationships between relatives at one locus. *PLoS One.* 8: e57003.

Gillois, M. 1964. *La relation d'identité en génétique*. Thesis, Faculté des Sciences de Paris.

Goldgar, D. E. 1990. Multipoint analysis of human quantitative genetic variation. *Am. J. Hum. Genet.* 47: 957–967.

Guo, S. W. 1994. Computation of identity by descent proportions shared by two siblings. *Am. J. Hum. Genet.* 54: 1104–1109.

Guo, S. W. 1995. Proportion of genome shared identical by descent by relatives: concept, computation, and applications. *Am. J. Hum Genet*. 56: 1468–1476.

Guo, S. W. 1996. Variation in genetic identity among relatives. *Hum. Hered.* 46: 61–70.

Haldane, J. B. S. 1919. The combination of linkage values and the calculation of distance between the loci of linked factors. *J. Genet.* 8: 299–309.

Haley, C. S., Knott, S. A. y Elsen, J. M. 1994. Mapping quantitative trait loci in crosses between outbred lines using least squares. *Genetics*. 136: 1195–1207.

Han, L. y Abney, M. 2011. Identity by descent estimation with dense genome-wide genotype data. *Genet. Epidemiol.* 35: 557–567.

Han, L., y Abney, M. 2013. Using identity by descent estimation with dense genotype data to detect positive selection. *Europ. J. Hum. Genet.*, 21(2): 205–211.

Harris, D. L. 1964. Genotypic covariance between inbred relatives. *Genetics*. 50: 1319−1348.

Henderson, C. R. 1984. *Applications of linear models in animal breeding*. University of Guelph, Guelph, Ontario, Canada.

Hill, W. G. 1993 a. Variation in genetic identity within kinships. *Heredity.* 71: 652−653.

Hill, W. G. 1993 b. Variation in genetic composition in backcrossing programs. *J. Hered*. 84: 212–213.

Hill, W. G. y Weir, B. S. 2011. Variation in actual relationship as a consequence of mendelian sampling and linkage. *Genet. Res.* 93: 47–64.

Hill, W. G. y Weir, B. S. 2012. Variation in actual relationship among descendants of inbred individuals. *Genet. Res*. 94: 267–274.

Jacquard, A. 1974. *The genetic structure of populations*. Springer-Verlag, New York, USA.

Karigl, G. 1981. A recursive algorithm for the calculation of identity coefficients. *Ann. Hum. Genet.* 45: 299–305.

Kenward, M. G. y Roger J. H. 1997. Small sample inference for fixed effects from restricted maximum likelihood. *Biometrics*. 53: 983–997.

Kong, A., Gudbjartsson, D. F., Sainz, J., Jonsdottir, G. M., Gudjonsson, S. A., Richardsson, B., Sigurdardottir, S., Bernard, J., Hallbeck, B., Masson, G., Shlien, A., Palsson, S. T., Frigge, M. L., Thorgeirsson, T. E., Gulcher, J. R. y Stefansson, K. 2002. A high-resolution recombination map of the human genome. *Nat. Genet.* 31: 241–247.

Lander, E. S. y Green P. 1987. Construction of multilocus genetic linkage maps in humans. En *Proceedings of the National Academy of Sciences of the United States of America*. 84: 2363–2367.

Lander, E. S. y Botstein, D. 1989. Mapping mendelian factors underlying quantitative traits using RFLP linkage maps. *Genetics*. 121: 185−199.

Leutenegger, A. L., Prum, B., Génin, E., Verny, C., Lemainque, A., Clerget-Darpoux, F. y Thompson, E. A. 2003. Estimation of the inbreeding coefficient through use of genomic data. *Am. J. Hum. Genet*. **7**3: 516**–**523.

Li, X., Yin, X. y Li, J. 2010. Efficient identification of identical-by-descent status in pedigrees with many untyped individuals. *Bioinformatics* . 26: 191−198.

Liu, J. M., Jansen, G. B. y Lin, C. Y. 2002. The covariance between relatives conditional on genetic markers. *Genet. Sel. Evol*. 34: 657−678.

Ma, L., O'Connell, J. R., VanRaden, P. M., Shen, B., Padhi, A., Sun, C., Bickhart, D. M., Cole, J. B., Null, D. J., Liu, G. E., Da, Y. y Wiggans, G. R. 2015. Cattle sex-specific recombination and genetic control from a large pedigree analysis. *PLoS Genet*. 11: e1005387.

MacLeod, A. K., Haley, C. S., Woolliams, J. A. y Stam, P. 2005. Marker densities and the mapping of ancestral junctions. *Genet. Res.* 85: 69–79.

Malecot, G. 1969. *The mathematics of heredity*. W. H. Freeman, San Francisco, USA.

McPeek, M. S. y Sun, L. 2000. Statistical tests for detection of misspecified relationships by use of genome-screen data. *Am. J. Hum. Genet*. 66: 1076–1094.

Meuwissen, T. H., Hayes, B. J. y Goddard, M. E. 2001. Prediction of total genetic value using genome-wide dense marker maps. *Genetics*. 157: 1819–1829.

Nettelblad, C., Holmgren, S., Crooks, L. y Carlborg, O. 2009. cnF2freq: Efficient determination of genotype and haplotype probabilities in outbred populations using Markov models. *BICoB.* 307–319.

Pong-Wong, R., George, A. W., Woolliams, J. A. y Haley, C. S. 2002. A simple and rapid method for calculating identity-by-descent matrices using multiple markers. *Genet. Sel. Evol.* 33: 453–471.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A. R., Bender, D., Maller, J., Sklar, P., de Bakker, P. I. W., Daly, M. J. y Sham, P. C. 2007. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet*. 81: 559–575.

Rabiner, L. R. 1989. A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*. 77: 257–286.

Ramos, A. M., Crooijmans, R. P. M. A., Affara, N. A., Amaral, A. J., Archibald, A. L., Beever, J. E., Bendixen, C., Churcher, C., Clark, R., Dehais, P., Hansen, M. S., Hedegaard, J., Hu, Z. L., Kerstens, H. H., Law, A. S., Megens, H. J., Milan, D., Nonneman, D. J., Rohrer, G. A., Rothschild, M. F., Smith, T. P. L., Schnabel, R. D., Van Tassell, C. P., Taylor, J. F., Wiedmann, R. T., Schook, L. B. y Groenen, M. A. M. 2009. Design of a high density SNP genotyping assay in the pig using SNPs identified and characterized by next generation sequencing technology. *PLoS ONE*. 4: e6524.

Rasmuson, M. 1993. Variation in genetic identity within kinships. *Heredity*. 70: 266–268.

Risch, N. y Lange, K. 1979. Application of a recombination model in calculating the variance of sib pair genetic identity. *Ann. Hum. Genet*. 43: 177–186.

Stam, P., y Zeven, A. C. 1981. The theoretical proportion of the donor genome in near-isogenic lines of self-fertilizers bred by backcrossing. *Euphytica*, 30(2): 227-238.

Stefanov, V. T. 2002. Statistics on continuous IBD data: Exact distribution evaluation for a pair of full (half)-sibs and a pair of a (great-) grandchild with a (great-) grandparent. *BMC Genet.* 3: 7.

Suarez, B. K., Reich, T. y Fishman P. M. 1979. Variability in sib pair genetic identity. *Hum. Hered*. 29: 37–41.

Thompson, E. A. 1975. The estimation of pairwise relationships. *Ann. Hum. Genet*. 39: 173–188.

Thompson, E. A. 1993. Genetic importance and genomic descent. En *Population management for survival and recovery*. University of Chicago Press, Chicago, USA.

Thompson, E. A. 2000. *Statistical inference from genetic data on pedigrees*. NSF-CBMS Regional conference series in probability and statistics. Volumen 6. USA.

Thompson, E. A. 2007. The IBD process along four chromosomes. *Theoret. Pop. Biol*.. 73: 369–373.

Thompson E. A. 2013. Identity by descent: variation in meiosis, across genomes, and in populations. *Genetics*. 194: 301–326.

Toro, M. A., García-Cortés, L. A. y Legarra, A. 2011. A note on the rationale for estimating genealogical coancestry from molecular markers. *Genet. Sel. Evol*. 43: 27.

Tortereau, F., Servin, B., Frantz, L., Megens H. J., Milan, D., Rohrer, G., Wiedmann, R., Beever, J., Archibald, A. L., Schook, L.B. y Groenen, M. A. M. 2012. A high density recombination map of the pig reveals a correlation between sex-specific recombination and GC content. *BMC Genomics*. 13: 586.

VanRaden, P. M. 2007. Genomic measures of relationship and inbreeding. *Interbull Bulletin*. 37: 33–36.

VanRaden, P. M. 2008. Efficient methods to compute genomic predictions. *J. Dairy Sc*. 91: 4414–4423.

Vela-Avitúa, S., Meuwissen, T. H. E., Luan, T. y Ødegård, J. 2015. Accuracy of genomic selection for a sib-evaluated trait using identity-by-state and identity-by-descent relationships. *Genet. Sel. Evol*. 47: 9.

Visscher, P. M., Medland, S. E., Ferreira, M. A. R., Morley, K. I., Zhu, G., Cornes, B. K., Montgomery, G. W. y Martin, N. G. 2006. Assumption-free estimation of heritability from genome-wide identity-by-descent sharing between full siblings. *PLoS Genet*. 2: e41.

Visscher, P. M. 2009. Whole genome approaches to quantitative genetics. *Genetica*, 136: 351–358.

Wakeley, J., King, L., Low, B. S. y Ramachandran, S. 2012. Gene genealogies within a fixed pedigree and the robustness of Kingman's coalescent. *Genetics*. 190: 1433-1445.